

UC Irvine

UC Irvine Previously Published Works

Title

SIPdb: a stable isotope probing database and analytical dashboard for linking amplicon sequences to microbial activity using reverse ecology

Permalink

<https://escholarship.org/uc/item/9nf8d16q>

Journal

ISME Communications, 6(1)

ISSN

2730-6151

Authors

Trentin, Alex Batista

Simpson, Abigayle

Kimbrel, Jeffrey A

et al.

Publication Date

2026-01-14

DOI

10.1093/ismeco/ycag216

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

SIPdb: a stable isotope probing database and analytical dashboard for linking amplicon sequences to microbial activity using reverse ecology

Alex Batista Trentin ¹, Abigayle Simpson ^{1,2}, Jeffrey A. Kimbrel ³, Steven J. Blazewicz ³, Roland C. Wilhelm ^{1,*}

¹Agronomy Department, Lilly Hall of Life Sciences, Purdue University, West Lafayette, IN 47907, United States

²Department of Ecology and Evolutionary Biology, Steinhaus Hall, University of California, Irvine, CA 92697, United States

³Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA 94550, United States

*Corresponding author: Agronomy Department, Purdue University, Lilly Hall of Life Sciences, 915 Mitch Daniels Blvd., West Lafayette, IN 47907, United States. E-mail: rcwilhelm@purdue.edu

Abstract

Stable isotope probing (SIP) connects microbial sequence data to diverse metabolic activities, but the lack of a unifying framework for SIP-derived data has limited its integration into broader strategies for ecological inference. Here, we introduce the SIPdb, an extensible SQLite database of curated nucleic acid SIP experiments (also in phyloseq format) paired with an interactive RShiny dashboard for analysis and visualization. The initial release compiles 22 studies covering 21 isotopolog substrates across diverse environments, standardized using the MISIP metadata standard. SIPdb provides a standardized pipeline accommodating the three most common SIP gradient fractionation strategies (binary, multi-fraction, and density-resolved), two incorporator designation strategies (fixed- and sliding-window), and four complementary differential abundance methods (DESeq2, edgeR, limma-voom, and ALDEx2). Using this pipeline, we identified over 42 000 unique amplicon sequence variants as isotope incorporators across 62 phyla. Benchmarking with SIPSim-generated synthetic datasets showed that position-resolved designs performed best and that differential abundance method contributed comparably to variation in incorporator designation, with DESeq2 suggested as a balanced default. Validation against original publications showed that, on average, SIPdb recovered 70.1% of author-reported incorporators, with discrepancies arising from differences in phylotyping or classification approaches. Finally, reanalysis of a non-SIP study of 1,4-dioxane degradation showed how SIPdb can both validate known degraders and uncover additional candidate taxa involved in community metabolism. SIPdb establishes a scalable platform for reverse ecology, enabling hypothesis generation, cross-study meta-analysis, and linking taxa to metabolic processes, while serving as an open, extensible resource to accelerate ecological interpretation in microbiome research.

Keywords stable isotope probing, reverse ecology, isotope incorporators, ecological inference gap

Introduction

Although the DNA sequencing revolution has significantly advanced our understanding of microbial diversity, efforts to interpret the ecological roles of community members remain limited by a persistent gap between sequence-based surveys and verified ecological function, which we refer to here as the ecological inference gap [1, 2]. This gap reflects a fundamental mismatch between commonly used sequencing methodologies and the inherent ecological and evolutionary properties of microbial systems, including variable gene copy numbers, strain-level heterogeneity, and the sparse availability of directly observed functions and activities [3–6]. A major contributor to this gap is the reliance on culture-based characterizations as a proxy for ecological function, and the use of taxonomic identities to link to these established bodies of ecological knowledge. This approach is limited by the fact that many taxa detected through high-throughput

sequencing lack experimentally verified ecological or functional traits, and by the ambiguity that results from reclassifications of taxonomic groups over time, including a recent revision of prokaryote nomenclature [7]. Resources are needed that move beyond dependence on individual studies and literature review, allowing ecological interpretations and assumptions about microbiome patterns to be tested directly. In this context, reverse ecology offers a promising approach that centers on the relationships between genetic sequences, their environmental distributions, and their *in situ* functional associations.

Reverse ecology is an approach that infers the ecological roles and traits of microbes from observed patterns in environmental sequence data, such as the distribution of genes, genomes, or taxa, across environmental gradients, host associations, or experimental conditions [8, 9]. Rather than focusing on characteristics of an organism, reverse ecology begins with a context, such as a soil disturbance or host,

Received: 4 March 2026. **Revised:** 14 June 2026. **Accepted:** 22 July 2026

© The Author(s) 2026. Published by Oxford University Press on behalf of the International Society for Microbial Ecology.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

and identifies the genes, transcripts, operons, regulatory elements, or genomes that are consistently associated with it. This strategy leverages the power of DNA sequencing not just to detect presence, but to identify environment-wide associations and ecological signals in large volumes of microbiome data. Recent studies have used reverse ecology to highlight the roles of rare microbes in adaptation to diverse abiotic factors [10] and functional genes in relation to soil properties [11], and to associate specific bacterial taxa with responses to tillage and irrigation regimes in agricultural soils [12]. This approach is particularly well suited to pattern detection methods, such as machine learning [13], and gains further power when paired with experimental strategies like stable isotope probing (SIP), which directly link sequence data to in situ metabolic activity [9, 14]. SIP functions as a distinct, activity-based form of reverse ecology in which the inference of ecological role is grounded not in distributional patterns alone, but in direct experimental evidence of substrate assimilation, linking sequence identity to metabolic activity under environmentally relevant conditions. Overall, the effectiveness and scalability of reverse ecology is enhanced by the availability of curated, context-rich sequence databases [15–18], resources that may support training of predictive algorithms.

Nucleic acid SIP provides a powerful experimental complement to reverse ecology by directly linking specific microbial sequences with functional activities in natural contexts [19, 20]. In a SIP experiment, environmental samples (e.g. soils, sediments, or hosts) are supplied with an isotopically labeled “heavy” substrate (e.g. ^{13}C , ^{15}N , or ^{18}O -enriched isotopologs). Organisms that metabolize the labeled substrates, or their by-products (i.e. “cross-feeding”), incorporate the heavy isotopes into their nucleic acids, which can then be separated based on their buoyant density shift and sequenced to designate “incorporators”: amplicon sequence variants (ASVs) or taxa [21–25]. This technique circumvents the need for inferring culture-based or taxonomic traits and instead provides a direct link to metabolic activity, though caveats do apply (see Current Limitations section). SIP experiments have revealed the ecological roles of previously uncultivated microbes in processes such as degradation of complex organic matter [26], microbial growth and turnover [27, 28], and trophic interactions like predation [22, 29]. The cross-study reuse has underscored the potential of aggregated SIP data to reveal broader ecological patterns [29]. Our research demonstrates how a curated database of nucleic acid SIP sequence data (“SIPdb”) can enhance the efficacy of a reverse ecology approach by linking microbial taxa to isotope incorporation patterns that reflect metabolic activity and ecological associations under environmentally relevant conditions.

Here, we describe the development and validation of the SIPdb, a comprehensive and extensible database designed to synthesize SIP-derived amplicon data for designating isotope incorporators across diverse substrate isotopolog experiments. The SIPdb accommodates three primary types of nucleic acid SIP gradient fractionation strategies: simple binary (“heavy” vs. “light”), position-resolved, and density-resolved; and two isotope incorporator designation strategies: fixed- and sliding-window. In curating the SIPdb, we adopted the Minimum Information for any Stable Isotope Probing Sequence (MISIP) standard [30], ensuring future SIP data can be easily integrated. While the current release is based on V4-region 16S rRNA gene data, the SIPdb framework can be extended to accommodate full-length amplicons and whole-metagenome SIP datasets, allowing for greater genomic resolution as more sequencing data become available. To accompany the SIPdb, we developed an interactive RShiny dashboard

(www.sip-db.com) that enables researchers to explore, compare, and analyze their own microbiome data against curated SIP datasets. The dashboard supports users to perform sequence homology and taxonomy-based queries, along with data visualization. While SIPdb is not a quantitative SIP (qSIP) resource and does not provide absolute incorporation rates [31], it accommodates the most commonly generated nucleic acid SIP data types. Most importantly, it was designed to extend ecological interpretation to non-SIP microbiome datasets, offering a scalable and empirically grounded framework for inferring putative ecological associations and metabolic participation from SIP-derived sequence data, as demonstrated in our study.

Materials and methods

Obtaining and curating SIP amplicon sequencing datasets

The initial release of SIPdb integrates data from 22 SIP studies encompassing 21 amended isotopologs and 14 distinct environmental biomes (Table S1). Metadata from all included studies were standardized following the MISIP framework (Table S2), ensuring interoperability and facilitating cross-study integration. In total, the database comprises 4770 amplicon sequencing libraries and 163 417 unique ASVs, of which 46 484 were designated as significant isotope incorporators in one or more datasets (including H_2^{18}O) under the most relaxed criteria ($\log_2\text{FC} > 1$, $p_{\text{adj}} \leq 0.05$). At present, only amplicon data overlapping the full V4 region of the 16S rRNA gene were kept, bounded by the primer set 515F (5'-GTGYCAGCMGCCGCGTAA-3') and 806R (5'-GGACTACHVGGGTWTCTAAT-3') [32]. Raw sequence data, from each study, were processed independently in QIIME2 (v. 2024.10) [33] and taxonomic classification was assigned using the SILVA 138.2 database [34]. The ASV table, taxonomic assignments, representative sequences, sample metadata, and phylogenetic tree for each study were imported into R (v4.4.1) [35] and formatted into a phyloseq object (v1.48.0) [36]. Detailed information regarding data acquisition from the NCBI Short Read Archive, the specific parameters of the QIIME2 and DADA2 processing workflows, and the adaptive thresholds used for sparsity and quality filtering is available in the Supplementary Methods.

Approaches to SIP incorporator designation

We developed a largely automated pipeline to standardize the identification of isotope incorporators from nucleic acid SIP datasets based on the differential abundance of sequences, either ASVs or taxa, in nucleic acid pools from isotopically enriched (“labeled”) and natural abundance (“unlabeled”) samples. The pipeline is coded in R and accepts phyloseq objects containing (i) an ASV count table, (ii) corresponding taxonomic assignments, and (iii) MISIP-formatted metadata including any other study factors. Importantly, the incorporator-designation pipeline does not use the phylogenetic tree or the representative ASV sequences; these elements are included in the phyloseq object only to support downstream analyses (e.g. UniFrac-based diversity metrics) and are not involved in determining incorporator status. Using SIP metadata and the count table, isotope incorporators are designated based on one of three gradient fraction cases and two window approaches for *in silico* combining of fractions. The pipeline outputs standardized results for each dataset, including: (i) taxon-level statistics for differential abundance ($\log_2$ fold-change, p -values, adjusted p -values) generated by multiple methods (DESeq2 [37],

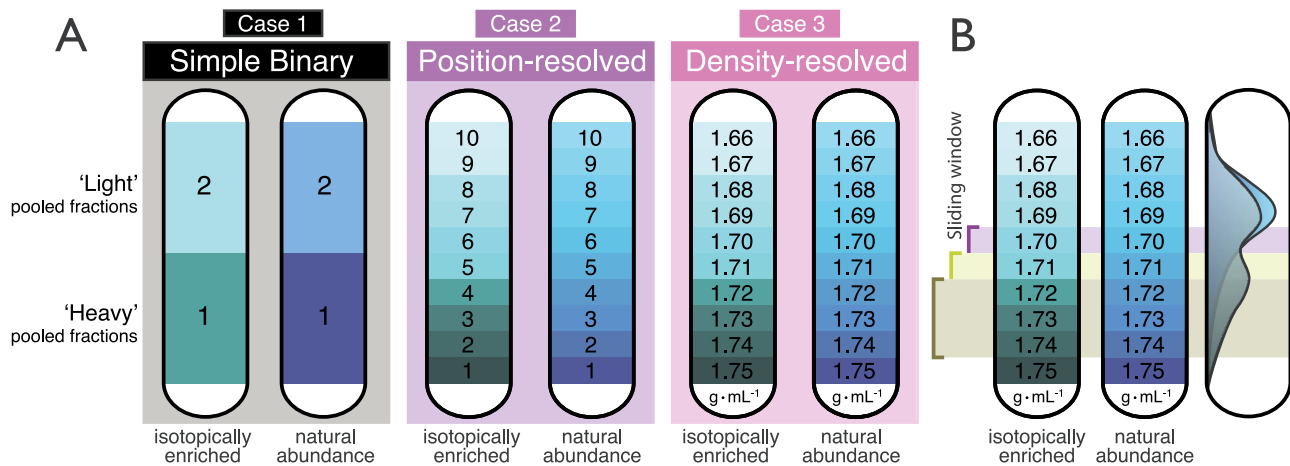


Figure 1 SIP gradient fractionation and windowing strategies used to designate isotope incorporators. In (A), the three SIP study types supported by SIPdb: Case 1 (simple binary comparison of pooled “light” and “heavy” fractions); Case 2 (position-resolved gradients with discrete fraction numbers); and Case 3 (density-resolved gradients with measured buoyant densities). In (B), the sliding-window approach applied to position- and density-resolved datasets, in which overlapping sets of adjacent gradient positions are compared between labeled and unlabeled samples.

edgeR [38], limma-voom [39], and ALDEx2 [40]), (ii) corresponding metadata describing incorporator designation strategy (heavy fraction, sliding window, or substrate-specific comparisons), and (iii) complete taxonomic and study-level provenance information. These results are compiled into a master table and summary reports that are subsequently imported into the SQLite SIPdb database (see next section). All code to reproduce the database is available on GitHub: <https://github.com/alexbtrentin/sipdb-pipeline>.

To adapt to variation in SIP gradient fractionation approaches, the pipeline accommodates three different gradient structures when designating isotope incorporators: (case 1) “binary”, where fractions are pooled into heavy and light categories; (case 2) “position-resolved”, with multiple sequential fractions but no buoyant density information; and (case 3) “density-resolved”, where fractions include buoyant density data (Fig. 1A). To flexibly identify incorporators, the pipeline integrates four established differential abundance methods (DESeq2, edgeR, limma-voom, and ALDEx2) and applies two complementary strategies for contrasting differential abundance between labeled and unlabeled controls to designate incorporators: a fixed-window and sliding-window approach, as previously described [41].

Pooling fractions (*in silico*) along the SIP gradient can improve sensitivity for detecting isotope incorporators, particularly when individual fractions have low sequencing depth. Pooling is also essential for case 2 studies, where density information is lacking and corresponding positions may differ slightly in buoyant density. By aggregating adjacent gradient positions, contrasts between labeled and unlabeled samples become less affected by stochastic variation and yield more reliable statistical inference. We implemented two window-based pooling strategies: a fixed-window approach, which designates the heaviest fraction range based on a consistent proportion of available positions within each study, and a sliding-window approach, which sequentially compares sets of adjacent gradient positions across the full gradient (Fig. 1B). We validated and compared fixed-window, sliding-window, and adaptive k-nearest-neighbor approaches across diverse datasets to determine the most reliable strategy for identifying isotopically labeled taxa with full details on these validation procedures and statistical calculations provided in the Supplementary Methods.

Validation of isotope incorporator designation with synthetic microbiome data

We evaluated the sensitivity, precision, specificity, and overall accuracy of incorporator designation using synthetic SIP datasets generated with SIPSIM [42]. One hundred independent SIPSIM simulations were generated using genome-derived GC-content distributions and defined isotope-incorporation status. Each simulation ($n = 100$ datasets) was analyzed using all three gradient fractionation strategies: binary, position-resolved, and density-resolved. Synthetic communities varied in richness, with 100, 150, or 219 taxa per simulation, and in the proportion of true incorporators, with 2%, 5%, 10%, 15%, 18%, or 30% of taxa designated as isotope incorporators. Isotope incorporation intensity was varied across simulations using atom-percent enrichment levels of 25%, 50%, 75%, or 100%. Each simulation included four isotope-labeled libraries and four natural-abundance control libraries. Their expected distribution across CsCl buoyant-density gradients was modeled with SIPSIM, and the resulting fraction-level count tables were subjected to simulated PCR amplification and sequencing-depth subsampling. Ground truth was validated by confirming that all designated incorporators shifted in all labeled libraries and that no non-incorporator taxa exhibited isotope-induced buoyant-density shifts. Each synthetic dataset was processed through our complete analysis pipeline to assess the sensitivity, precision, specificity and F1 score (details in Supplementary Methods).

Validation against original SIP studies

To further evaluate the accuracy of SIPdb, we compared the designated incorporator results from our pipeline against those reported in the original study describing each dataset ($n = 22$ studies). For all included studies, we systematically reviewed the corresponding manuscripts and supplementary materials to extract the primary isotope-incorporating taxa reported by the authors. When available, ASV-level sequence identifiers were directly retrieved. Otherwise, genus- or family-level incorporators were compiled from tables, figures, or text descriptions. These author-reported incorporators were assembled into a curated reference list spanning 411 taxa across 21

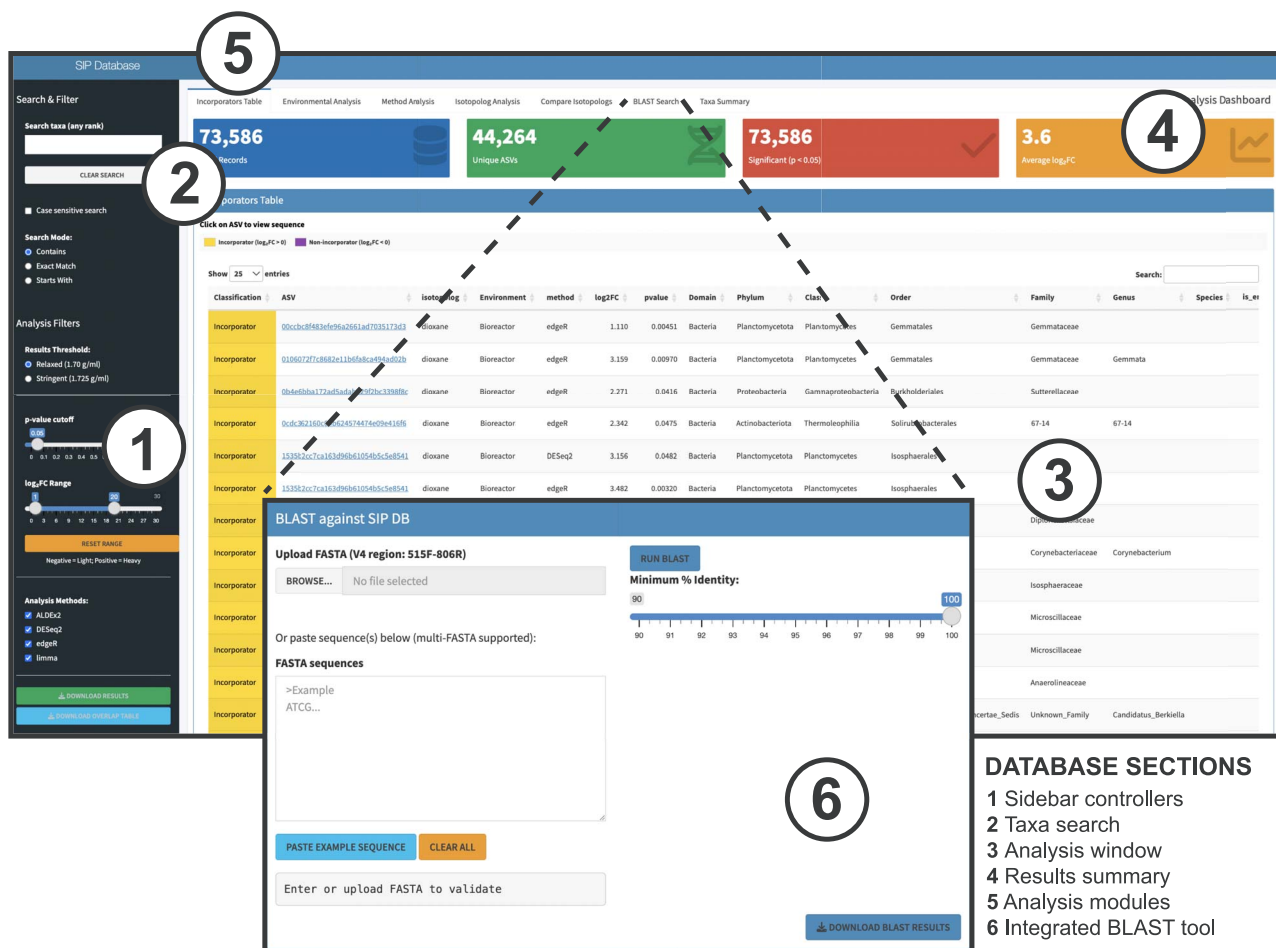


Figure 2 The SIPdb analysis dashboard. Screenshot of the interactive RShiny web application (www.sip-db.com) developed for querying and visualizing SIPdb data. Numbered components correspond to the labels printed on the figure: (1) sidebar controls for filtering results by density threshold (relaxed vs. stringent), p -value and $\log_2FC$ cutoffs, and differential abundance method; (2) taxa search field for querying records by taxonomy at any rank, including ASV ID; (3) analysis window displaying the incorporators table, in which each ASV is shown with its differential abundance statistics ($\log_2FC$, p -value), taxonomic classification, isotopolog, environment, and method; (4) summary panels showing study and dataset metrics; (5) analysis modules navigation bar providing access to the dashboard's seven modules; and (6) integrated BLAST tool for matching user-provided 16S rRNA gene V4 sequences (515F–806R) against SIPdb incorporator sequences. Together, these components enable interactive exploration of isotope incorporation data and direct annotation of users' own microbiome datasets.

isotopologs, which served as the basis for validation against SIPdb results (Table S3).

We then queried SIPdb to determine whether each reported ASV/taxon was detected as a significant incorporator ($p_{adj} \leq 0.05$, $\log_2FC > 1$) using any of the four differential abundance methods, given the variability of approaches used among studies. A taxon was considered “validated” if it appeared in the SIPdb results for the corresponding isotopolog and study, regardless of which method(s) detected it (Table S3). For validated taxa, we recorded which methods identified the incorporator and the mean $\log_2FC$ value. When available, ASV-level identifiers or representative sequences reported in the original publications were directly used for validation against SIPdb results. When ASV-level information was not provided, validation was performed at the taxonomic rank reported by the authors (typically genus or family). A reported taxon was considered validated if it appeared as a significant incorporator in SIPdb for the corresponding study and isotopolog using any differential abundance method. This mixed-resolution validation reflects reporting practices in the original studies and avoids excluding datasets lacking ASV-level identifiers.

Constructing the SIPdb relational database

For the initial release of SIPdb, we compiled the SIP results into a relational SQLite database that links differential abundance outcomes to study metadata, taxonomic lineages, and representative sequences. Incorporators were designated using both relaxed ($\geq 1.70 \text{ g}\cdot\text{mL}^{-1}$) and stringent ($\geq 1.725 \text{ g}\cdot\text{mL}^{-1}$) density thresholds to provide flexible sensitivity levels, with consensus calls derived from the integration of multiple statistical frameworks and windowing strategies. To ensure data integrity, we implemented a rigorous validation pipeline and integrated a curated BLAST database for sequence homology-based queries. Comprehensive details regarding the database architecture, filtering criteria, and validation protocols are provided in the Supplementary Methods.

SIPdb analysis dashboard

To facilitate interactive exploration and broad accessibility, we developed a comprehensive Shiny web [43] application that serves as the analysis dashboard for SIPdb (Fig. 2), which is available to users at www.sip-db.com. The dashboard allows users to explore isotope

incorporators across diverse isotopolog substrates and environmental contexts, and to annotate their own microbiome datasets using a web-based BLAST search against representative sequences of incorporator ASVs in SIPdb. Users can adjust statistical thresholds (p -value cutoff, $\log_2\text{FC}$ ranges, relaxed vs. stringent density thresholds) and compare differential abundance methods. The SIPdb analysis workflow (phyloseq objects and R scripts), the SQLite database and sequence data, and the dashboard app (“app.R”) are all open source and available for local use or adapting for future uses cases via GitHub (<https://github.com/alexxtrentin/sipdb-dashboard>). The dashboard comprises seven analysis modules: *Incorporators Table*, *Environmental Analysis*, *Method Analysis*, *Isotopolog Analysis*, *Compare Isotopologs*, *Taxa Summary*, and *BLAST Search*, which enable users to visualize community patterns, compare substrate-specific enrichment, and annotate sequences against the database. Comprehensive descriptions of module features are provided in the Supplementary Methods.

Ecological inference with SIPdb

SIPdb can be used in two primary ways to derive ecological insights: (i) conducting meta-analyses across SIP datasets and (ii) mapping patterns of isotope incorporation onto non-SIP microbiome data to infer potential ecological associations or metabolic roles. To illustrate the first case, we visualized how substrate chemical properties influence microbial incorporation patterns across SIPdb by performing principal component analysis integrating substrate chemistry with incorporation metrics. For each substrate, we compiled six variables: (i) three chemical descriptors (solubility, hydrophobicity [xLogP3], and molecular weight from PubChem [44], Table S4) and (ii) three incorporation metrics from significant enrichment results ($p_{\text{adj}} < 0.05$): number of incorporating taxa (unique ASVs/OTUs), mean differential abundance in isotopically-enriched DNA or “effect size” ($\log_2\text{FC}$), and taxonomic breadth (number of unique phyla, a proxy for metabolic generality). Polymeric substrates (e.g. cellulose, DNA, chitin) were excluded because molecular descriptors are undefined for macromolecules with variable chain lengths, and substrates with incomplete chemical annotation were also excluded, yielding a total of 14 substrates in our analysis. Count-based variables were $\log_{10}$ -transformed to address right-skewed distributions, and PCA was performed on scaled and centered data using `prcomp` in R [35]. To examine relationships between substrate chemistry and labeling strength, substrates were classified into effect-size categories (Low, Medium, High) based on tertiles of mean $\log_2\text{FC}$, and differences in PC1 scores among categories were assessed using Kruskal-Wallis tests with Dunn’s post-hoc comparisons (Benjamini-Hochberg correction).

To illustrate the second case, we evaluated whether SIPdb could aid interpretation of non-SIP microbiome datasets. We re-analysed 16S rRNA gene data from Li et al. (2025) [45], which examined two 1,4-dioxane-contaminated field sites with contrasting degradation outcomes (BioProject: PRJNA1183415). Data were processed using the same workflow applied to SIPdb studies. Briefly, raw reads were processed in QIIME2 (v. 2024.10) using DADA2 for denoising and chimera removal, yielding 13 176 ASVs, and taxonomy was assigned using SILVA 138.2 [34]. Putative 1,4-dioxane-responsive taxa were identified based on differential abundance (assessed using DESeq2, consistent with the analytical framework applied in the original study) between microcosms amended with 1,4-dioxane and unamended controls.

Differentially abundant ASVs were queried against SIPdb using the integrated BLAST tool; however, few met the criterion of 100% identity across the full V4 region. In contrast, many ASVs that were not differentially abundant exhibited perfect sequence matches to SIPdb records. Consequently, a taxonomy-based matching approach was adopted. Differentially abundant ASVs from Li et al. were aggregated at the genus level, and genera were considered SIPdb-associated if any ASV classified to that genus had a matching record in SIPdb. Genera were then classified into three categories: *validated* (reported by Li et al. and present in the SIPdb match set), *recovered* (not reported by Li et al. but present in SIPdb as dioxane or phenanthrene incorporators), or *absent* (enriched in contaminated samples but not represented in SIPdb). SIPdb matches associated with other substrates (e.g. methane, butane, acetate) were retained for interpretive context only. Statistical comparisons used Wilcoxon tests for relative abundance and Fisher’s exact tests for presence-absence.

Results and discussion

Overview of SIPdb incorporator designations

Version 1 of the SIPdb currently contains 22 studies, representing a broad range of environments (14 different environment types) and substrates (21 isotope labelled substrates). The database identified 46 484 unique ASVs as significant isotope incorporators across all individual differential abundance methods ($p_{\text{adj}} \leq 0.05$, $\log_2\text{FC} > 1$), of which 33 939 were retained as consensus incorporators (≥ 2 methods). The isotopolog substrates span a broad range of metabolic contexts, including simple carbon substrates (glucose, methane), fermentation products (acetate, lactate, oxalate), complex carbohydrates (alginate, cellulose), aromatic compounds (phenanthrene, vanillin), and specialized substrates such as alkanes (butane), xenobiotics (dioxane, sulfamethoxazole), and biomolecules (DNA, amino acids). Within this set, ^{18}O -labeled water occupies a distinct functional category, as it reflects the prevailing metabolic activity of all physiologically active organisms, independent of any specific substrate amendment. This allows ^{18}O -H₂O to serve as a baseline indicator of the actively growing microbiome within each environmental context, while underscoring the need to interpret SIPdb-derived incorporation patterns in light of the distinct ecological meanings associated with different isotopologs.

The SIPdb incorporator set spanned 62 phyla, with members of Pseudomonadota (27.6%), Bacteroidota (11.0%), and Chloroflexota (9.8%) representing the most frequently detected incorporators across studies. The predominance of Pseudomonadota as incorporators highlights the broad metabolic versatility and successful ecological strategies of this diverse lineage of bacteria, with members designated as incorporators across diverse substrates (Fig. 3). Bacteroidota and Chloroflexota were also major isotope incorporators, though with a higher degree of substrate-specific associations.

Single-carbon (C1) substrates yielded the highest numbers of designated incorporators, reflecting both the ubiquity of C1 metabolism and the foundational role of these organisms in supporting carbon assimilation at higher trophic levels. Bicarbonate was incorporated by 3633 ASVs across 41 phyla (mean $\log_2\text{FC} = 2.35$), indicating that the capacity to assimilate inorganic carbon is phylogenetically widespread, encompassing $\sim 7\%$ of ASVs detected in SIPdb. This proportion is higher than, but broadly comparable to, genome-based estimates suggesting that $\sim 2\%$ of bacterial and archaeal genomes encode

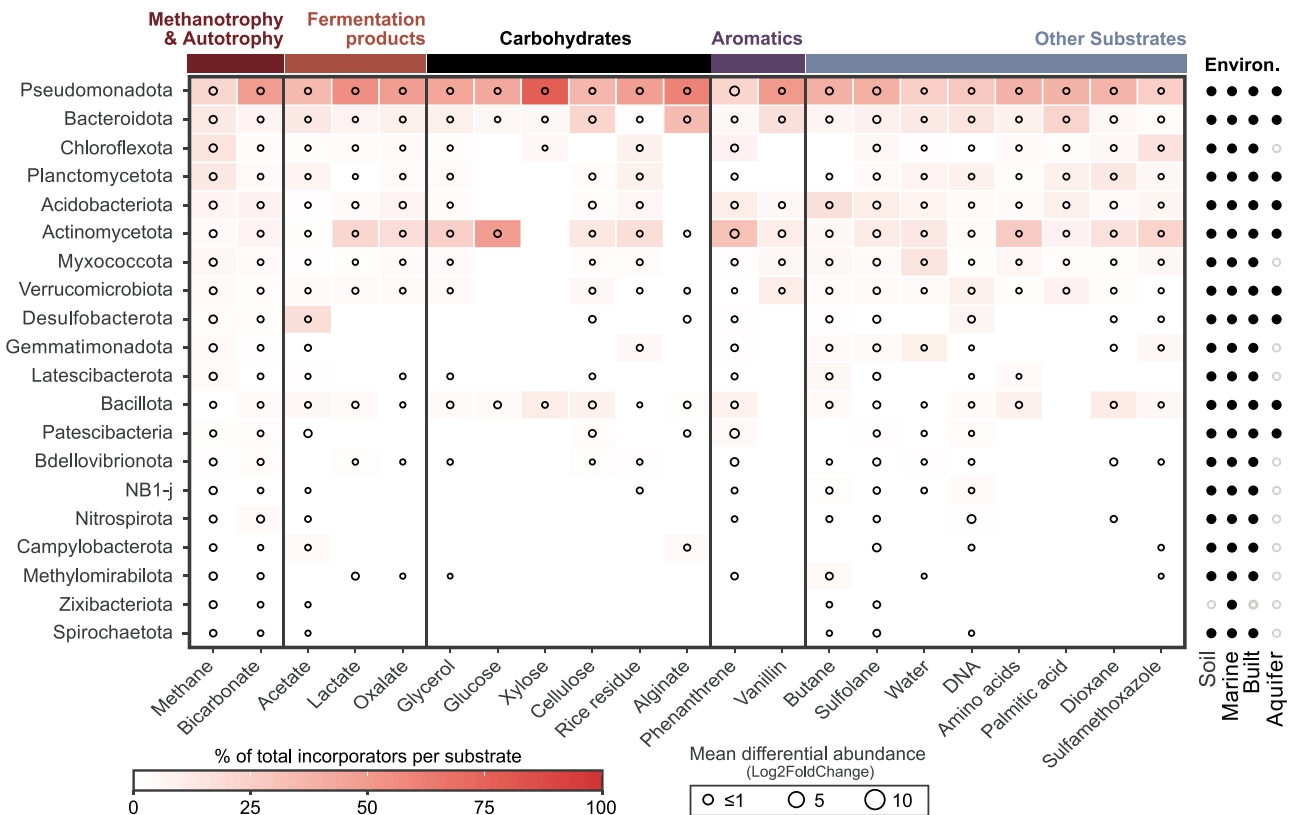


Figure 3 Distribution of isotope incorporators across substrates, phyla, and environments. Bubble heatmap summarizing isotope-incorporating ASVs for the 20 most prevalent phyla (rows) across the 21 labeled substrates (columns) in SIPdb. For each phylum–substrate combination, the square color shows the percentage of that substrate's total designated incorporators assigned to the phylum and circle size corresponds to the mean differential abundance ($\log_2FC$). The right panel indicates the environment types in which each phylum was detected as an incorporator (soil, marine, built, aquifer). Filled circles denote detection and open circles non-detection.

complete carbon fixation pathways [46]. Importantly, our result is likely an overestimate given it likely reflects a combination of primary autotrophy, mixotrophy, and downstream incorporation via cross-feeding and trophic transfer, processes that cannot be disentangled using SIP alone. Consistent with this interpretation, Chloroflexota exhibited strong methane incorporation (3624 ASVs), highlighting their potential importance within methanotrophic and methane-supported microbial communities [47]. Many methane studies in SIPdb spanned long incubation periods (24 to 720 h), which likely promote secondary consumption and food-web propagation of ^{13}C , explaining the large apparent taxonomic breadth of methane incorporators despite the taxonomically narrow set of known primary methanotrophs. In contrast to C1 compounds, a relatively narrow set of phyla were designated as glucose incorporators in SIPdb, which is unexpected given that glucose uptake is broadly distributed among bacteria [48]. These observations suggest that growth traits (uptake kinetics and specific growth rates), substrate concentration, incubation time, and bioavailability, not just the presence of metabolic pathways, determine whether a taxon is designated as an incorporator [44, 49].

Together, the contrasting breadth of incorporators observed for C1 substrates versus glucose highlights the ecological and methodological complexity inherent to SIP-derived incorporator data, emphasizing that incorporator breadth in SIP experiments reflects ecological context and trophic dynamics rather than simple substrate-pathway relationships.

Benchmarking incorporator designation with synthetic SIP datasets

Across 100 synthetic SIP datasets generated with SIPSIm, performance was shaped by both gradient fractionation strategy and differential abundance method. The position-resolved approach yielded the strongest mean F1 score (0.776), density-resolved was intermediate (0.724), and binary was the least effective in incorporator designation (0.701) (Fig. 4A and B). The binary strategy gave the most stringent results, with high precision (0.983) and specificity (0.999), recovering fewer than half of true incorporators (low sensitivity) while almost never producing a false call. The position-resolved strategy achieved the best balance of precision (0.822) and sensitivity (0.735), while the density-resolved strategy was intermediate (precision = 0.812; sensitivity = 0.604). The stronger performance of gradient-resolved strategies over binary pooling is consistent with prior benchmarking [42, 50]. Binary pooling carries the additional risk that natural variation in genome G + C content can shift the buoyant density of unlabeled DNA by up to $\sim 0.05 \text{ g mL}^{-1}$, allowing unlabeled to be miscalled as incorporated when position or density information is unavailable [41]. Resolved designs, which retain gradient structure, are less susceptible to this confounding.

Method choice and gradient design contributed comparably to variation (Fig. 4C, Table S5). Among the SIPSIm datasets for which all methods returned results, edgeR and DESeq2 performed best and had comparable scores, with edgeR ($F1 = 0.777$, precision 0.852, sensitivity 0.785) favoring sensitivity and DESeq2 precision ($F1 = 0.773$, precision

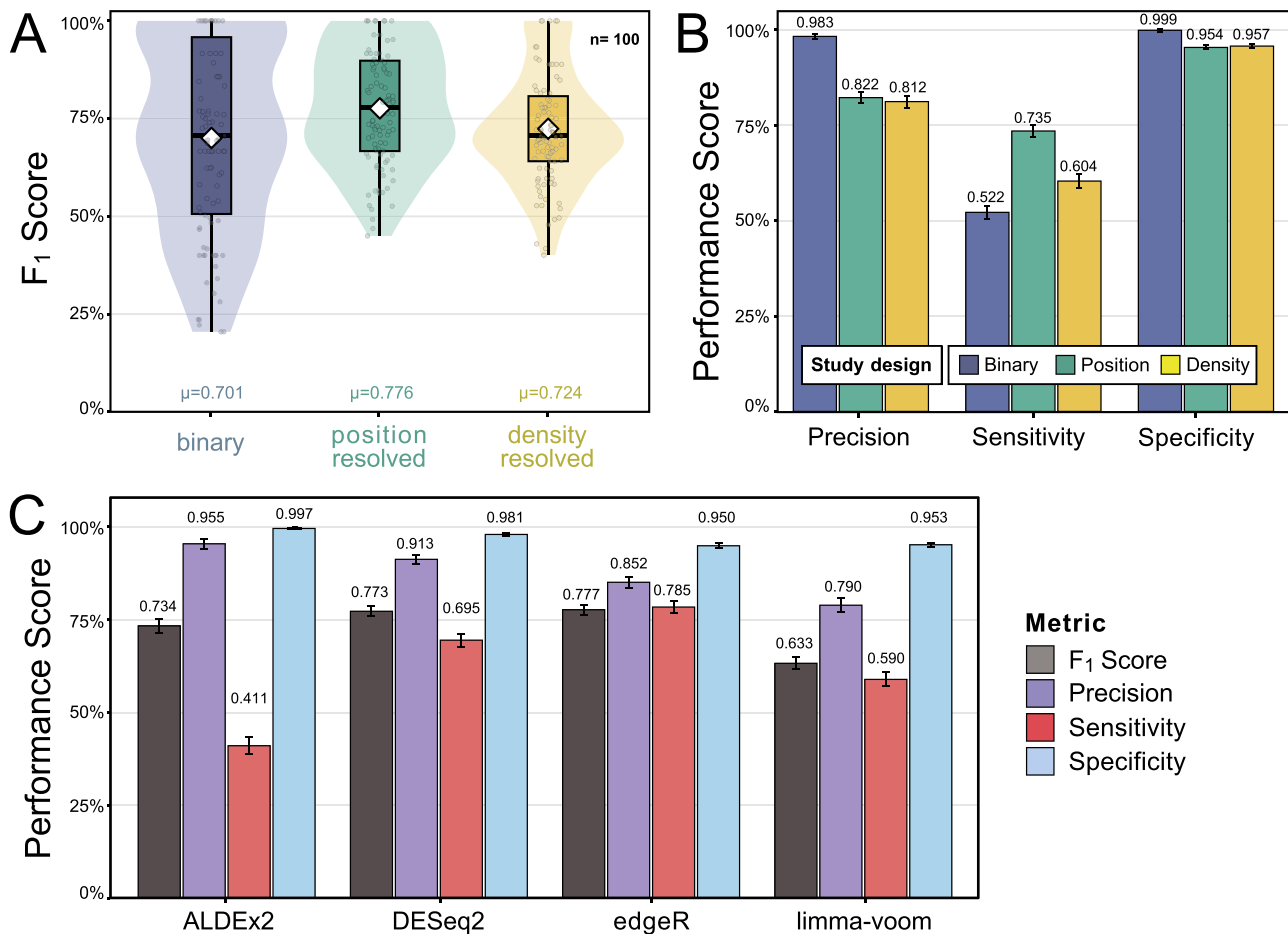


Figure 4 Benchmarking of gradient fractionation strategies and differential abundance methods using synthetic SIP datasets generated with SIPSIm. In (A), violin plots with overlaid box plots showing the distribution of F1 scores for incorporator designation under each gradient fractionation strategy (binary, position-resolved, density-resolved; $n = 100$ synthetic datasets each). Boxes show the median and interquartile range, whiskers extend to $1.5 \times$ the interquartile range, and white diamonds mark the mean. In (B), mean precision, sensitivity, and specificity for each gradient fractionation strategy, colored by strategy. Values are printed above each bar. In (C), mean F1 score, precision, sensitivity, and specificity for each differential abundance method (ALDEx2, DESeq2, edgeR, limma-voom). Error bars denote standard error. Across all panels, performance is evaluated against the known incorporator status of taxa in the synthetic communities using the fixed-window strategy.

0.913, sensitivity 0.695). ALDEx2 was the most precise (0.955) but least sensitive (0.411), giving an intermediate F1 (0.734), and limma-voom was weakest (F1 = 0.633). The methods also differed markedly in applicability, with DESeq2, edgeR, and limma-voom returning results for every dataset, whereas ALDEx2 generated results for only 67.7% overall (48% of binary, 67% density, and 75% position-resolved designs), reflecting a known limitation of its interquartile log-ratio normalization on sparse contrasts [40] (details in Supplementary Methods). We tested whether substituting the zero denominator on these contrasts could recover the missing calls: across the benchmark it recovered 360 additional true incorporators but added 406 false positives, with the cost concentrated in density-resolved designs, where precision fell from 0.74 to 0.47, while binary and position-resolved designs were little affected. We therefore retained the IQLR denominator for SIPdb version 1 results. No method or design was best across all performance metrics: resolved designs were most sensitive, edgeR and DESeq2 most accurate and broadly applicable, and ALDEx2 most precise but least widely usable.

With the current criteria ($p_{\text{adj}} < 0.05$, $\log_2\text{FC} > 1$), edgeR accounts for a majority of per-method incorporator calls for SIPdb study datasets. Two factors drive this: (i) edgeR was among the most sensitive methods in our benchmark (0.785) and (ii) other methods had

limitations that reduced outputs, specifically, where ALDEx2 was able to compute only 43 of 113 contrasts (38%) using the default IQLR denominator approach. The overrepresentation of edgeR results reflects differences in statistical stringency and applicability, meaning the aggregated results of all four methods accessible via the SIPdb are valid, but should be interpreted as a discovery set rather than a confirmed inventory. For new SIP data, DESeq2 is a reasonable default for balanced designation, edgeR for broad discovery, and ALDEx2 for stringent calls on feature-rich contrasts. For filtering within the current release, stringency matters more than method choice: the union returns 46 484 ASVs (edgeR-dominated), the consensus subset (≥ 2 methods) 34 939 ASVs (recommended for hypothesis generation), and the strictest filter (≥ 3 methods) 3006 ASVs, yielding the highest likelihood of isotope incorporation. These results align with reports that method choice is a major source of variance in microbiome analyses [51].

Benchmarking against the original SIP studies

To validate SIPdb results against published findings, we systematically extracted the primary incorporators reported by authors in each of the 22 original studies, resulting in 411 taxa spanning genus

to phylum level across 21 distinct isotopologs (Table S3). Of these author-identified incorporators, 70.1% (288/411 rank-agnostic-taxa) were successfully detected in SIPdb using our standardized analysis pipeline, demonstrating strong concordance between the database and original study conclusions. Detection rates varied substantially by substrate: perfect validation (100%) was achieved for alginate (8/8 reported incorporators recovered) and DNA (6/6), with high rates for water (88.9%), glucose (88.9%), sulfamethoxazole (88.9%), bicarbonate (85%), butane (83.3%), and methane (81.1%).

Lower detection rates were observed for palmitic acid (7.7%, 1/13) and xylose (8.3%, 1/12 reported incorporators), both of which were reported in a single study in the current database. To determine whether these results reflected differences in statistical stringency rather than analytical failure, we performed a sensitivity analysis by relaxing the $\log_2FC$ threshold to >0 while maintaining other criteria. Detection rates increased to 69.2% (9/13) for palmitic acid and 33.3% (4/12) for xylose, demonstrating that many author-reported incorporators produce detectable enrichment signals in SIPdb but with effect sizes below our conservative threshold.

Non-detection of author-reported incorporators (29.9%) likely stems from (i) statistical threshold differences between original studies and SIPdb's standardized $p_{adj} \leq 0.05$ criterion, (ii) taxonomic assignment variation across reference databases and algorithms, (iii) data processing heterogeneity in normalization and quality filtering, and (iv) biological factors including stochastic sampling of low-abundance taxa. The strong overall concordance (70.1%) demonstrates that SIPdb successfully reproduces most published SIP findings while maintaining conservative statistical rigor, supporting its utility as a reliable resource for comparative SIP analysis.

Cross-study exploration of substrate-associated incorporator profiles

To validate the ecological coherence of SIPdb and explore how substrate chemistry shapes incorporator communities, we performed a principal component analysis integrating substrate physicochemical properties with incorporation metrics across 14 substrates. As a first validation, we asked whether SIPdb could recapitulate the well-established expectation that small, soluble substrates support broader or more diverse microbial incorporator communities compared with large, hydrophobic, or recalcitrant substrates, which require specialized degradative machinery [52]. The second principal component (PC2, 26.5%) confirmed this expectation, with loadings associated with physicochemical accessibility separating small, highly soluble substrates (e.g. acetate, glycerol) from hydrophobic or high-molecular-weight compounds (e.g. palmitic acid, sulfamethoxazole). That SIPdb recovers this pattern from aggregated cross-study data provides confidence that the database faithfully captures substrate-driven ecological patterns (Fig. 5).

The more informative axis, however, was the first principal component (PC1, 54.5%), which exhibited strong loadings for the proportion of incorporator taxa (0.50), the total number of incorporators (0.50), and taxonomic breadth (0.48). Together, these metrics capture variation in trophic accessibility, reflecting the extent to which a substrate's isotopic label propagates through the microbial food web, from primary consumers to downstream cross-feeders [53, 54]. Substrate positions along PC1 corresponded to significant differences in labeling strength (approximated by $\log_2FC$ in relative abundance), where high-effect size substrates clustered at the high-trophic-accessibility end

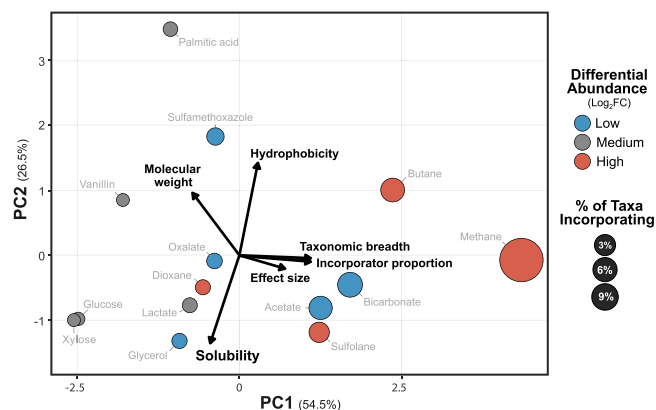


Figure 5 Principal component analysis relating substrate properties and incorporator patterns ($n = 14$ substrates). Biplot of a PCA in which each substrate is described by six variables: three physicochemical descriptors (solubility, hydrophobicity [$\times \log P_3$], molecular weight) and three incorporation metrics (effect size [mean $\log_2FC$], taxonomic breadth [number of unique phyla], and incorporator proportion [proportion of detected ASVs designated as incorporators]). Each point represents a labeled substrate, colored by mean effect size ($\log_2FC$) and scaled by the proportion of ASVs designated as incorporators for that substrate relative to the total number of ASVs detected in the corresponding study or studies. Arrows show the loadings of all six variables. Polymeric substrates (e.g., cellulose, DNA, chitin) and those with incomplete chemical annotation were excluded.

of PC1 (mean PC1 = 1.84), whereas medium-effect substrates occupied the low-trophic-accessibility region (mean PC1 = -1.73 ; Kruskal-Wallis $\chi^2 = 8.73$, $P = 0.013$). Notably, methane departs from this chemical expectation: despite being a small, non-polar gas, it is among the most broadly incorporated substrates in the database, reflecting the wide phylogenetic distribution of methane monooxygenase systems [55] rather than substrate lability. Together, these axes demonstrate that substrate identity structures microbial communities not only through chemical constraints on assimilation, but also through the extent to which isotopic label is accessed across trophic levels after primary assimilation, a pattern consistent with documented cross-feeding dynamics in SIP experiments [56–59].

To evaluate reproducibility across independent studies, we examined substrates tested multiple times. Mantel tests revealed a significant positive correlation between study replication and phylum-level similarity (Spearman $r = 0.256$, $P = 0.026$), indicating that substrates examined across diverse environments, including methane and cellulose, yielded more consistent incorporator communities than those from single studies. This pattern was not driven by differences in within-substrate variance (betadisper $P = 0.729$). Reproducibility was strongly scale-dependent, with phylum-level profiles exhibiting substantially higher overlap across studies (mean Jaccard similarity = 0.44 ± 0.21) than genus-level profiles (0.039 ± 0.05), an 11.3-fold difference. This scale dependence mirrors broader patterns in microbial ecology, in which high-level functional traits are conserved within major clades, while fine-scale taxonomic composition reflects local environmental filtering [60–62]. Practically, this suggests that SIPdb information is propagated across broad taxonomic ranks, with phylum- and class-level guilds providing more reliable guidance for identifying likely incorporators of novel substrates than genus-level assignments. Together, these results show that SIPdb captures generalizable ecological rules of substrate use that extend beyond primary metabolism to include trophic accessibility.

Reverse ecology approach using SIPdb

To evaluate whether SIPdb can inform mechanistic interpretation of non-SIP microbiome datasets, we mapped 16S rRNA gene profiles from a study of 1,4-dioxane degradation (Li et al., 2025) [45] onto SIPdb and assessed whether enriched taxa had prior isotope-based evidence of substrate assimilation. SIPdb currently contains one dioxane-degradation SIP dataset, conducted in activated sludge bioreactors, whereas Li et al. examined degradation in polluted groundwater sediments. Our goal was to assess whether SIPdb could help identify patterns in isotope-incorporating taxa, revealing genera with documented incorporation capacity that might otherwise be overlooked in site-specific analyses. Importantly, any matches we observe are likely to reflect both primary dioxane-degrading populations as well as cross-feeding taxa whose incorporation is driven by trophic propagation.

Exact OTU-level correspondence between datasets was rare, so a taxonomy-based approach was used to designate SIP incorporators. Of the 10 genera reported by Li et al. as enriched at dioxane-contaminated sites, five matched SIPdb records indicating incorporation by dioxane or the structurally related compound phenanthrene (Fig. 6). These validated taxa were not randomly distributed within the community; they included several of the most abundant populations, such as *Brevundimonas* (26.1% of total reads, $\log_2FC = 4.1$), *Rhodococcus* (25.1%, $\log_2FC = 2.7$), and *Paenarthrobacter* ($\log_2FC = 7.2$; 9.8%). This pattern suggests that even with limited dioxane-specific data, SIPdb can help identify ecologically dominant degraders relevant to bioremediation applications. The validation of *Brevundimonas* and *Streptococcus* through direct dioxane SIPdb matches, alongside *Rhodococcus*, *Paenarthrobacter*, and *Lysobacter* through phenanthrene records, demonstrates the value of warehousing structurally related compounds in the SIPdb as means for discovery. The inclusion of phenanthrene-incorporating taxa in our query reflects the observation that bacteria capable of PAH degradation often possess diverse catabolic capabilities and broad-substrate monooxygenase systems [63–65].

Five differentially abundant genera in Li et al. dioxane-amended microcosms were absent from SIPdb's dioxane/phenanthrene records: *Dietzia*, *Hydrogenophaga*, *Phyllobacterium*, *Geobacillus*, and *Flavobacterium*. These taxa played an important role in the study system, with *Dietzia* exhibiting the second-highest relative abundance increase in their experiment ($\log_2FC = 5.7$), while *Hydrogenophaga* was among the most abundant (13.6% of total reads). Notably, three of the absent genera, *Hydrogenophaga*, *Flavobacterium*, and *Geobacillus*, were identified as incorporators for methane or butane incorporation in the SIPdb. This observation has important mechanistic implications: soluble di-iron monooxygenases, including propane monooxygenases and butane monooxygenases, exhibit broad substrate specificity and are well-documented for the cometabolic oxidation of dioxane [66, 67]. Taxa capable of C2–C4 alkane oxidation may therefore possess latent dioxane degradation capacity through these promiscuous enzyme systems [68]. This result demonstrates how SIPdb may extend interpretation of results, enabling hypothesis generation beyond simple substrate matching. Querying taxa enriched for related pathways (e.g. C2–C4 alkane oxidation) allows users to infer cometabolic potential even when direct SIP evidence is unavailable. Regardless, users may anticipate that phylotypes and taxa key to their study system may be absent from SIPdb due to either gaps in database coverage or geographical and ecological differences.

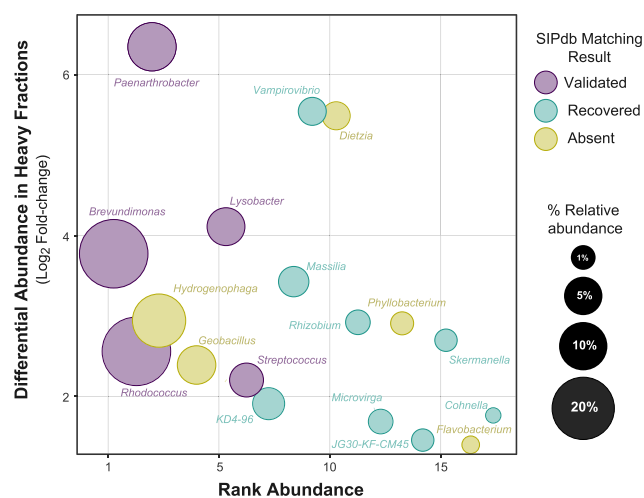


Figure 6 Reverse-ecology annotation of 1,4-dioxane-responsive taxa using SIPdb. Genera enriched in the 1,4-dioxane-amended microcosms of Li et al. (2025), a non-SIP dataset, plotted by abundance rank (x-axis; rank 1 = highest relative abundance) and differential abundance (y-axis; $\log_2FC$, dioxane-amended vs. unamended controls). Dot size indicates each genus's mean relative abundance in the contaminated samples, and color indicates its SIPdb matching result: “validated” (reported by Li et al. as dioxane-responsive and matching a SIPdb record), “recovered” (enriched but not reported by Li et al., yet matching a SIPdb record), and “absent” (enriched but lacking a matching SIPdb record). Because exact OTU-level correspondence between datasets was rare, matching was performed at the genus level.

Beyond confirming known degraders, SIPdb matches identified eight genera with documented dioxane and phenanthrene incorporation that were enriched at other contaminated sites but not reported in Li et al. Recovered genera included *Vampirovibrio* ($\log_2FC = 5.8$), *Massilia* ($\log_2FC = 3.5$), and multiple *Rhizobiales* members, including *Rhizobium*, *Skermanella*, and *Microvirga*. The presence of multiple *Rhizobiales* is notable given this order's known metabolic versatility [69, 70]. *Vampirovibrio*, a predatory bacterium, may represent indirect enrichment through predation on primary degraders rather than direct dioxane metabolism. This fact illustrates the importance of interpreting matches to the SIPdb with caution, as the SIPdb can reveal both direct incorporators and taxa responding indirectly to contaminant exposure [29, 71]. Our reverse ecology approach demonstrates the use of SIPdb to drive hypothesis generation.

Future advances and the promise of the SIPdb

Despite the relatively limited number of publicly available datasets in the SIPdb, due to legacy issues in data curation [30], our experimentation with the initial release of SIPdb revealed patterns in microbial substrate incorporation and demonstrated its promise as a tool for reverse ecology. All SIPdb entries are formatted according to the MISIP standard [30], ensuring its forward compatibility with newly generated SIP datasets. MISIP is slated for adoption as a formal MixS checklist, which will guide SIP data archival for integration in the SIPdb. Furthermore, the volume of SIP sequencing data is expected to grow as recent innovations in high-throughput, semi-automated workflows reduce the labor burden associated with SIP, cutting hands-on time by six-fold [56]. Previous advances in SIP computational methods were integral to the development of the SIPdb (e.g. HTS-SIP [50]) and continued progress in this space, including the use of synthetic

DNA internal standards for improved comparability and accuracy [72], will enhance SIPdb's utility. As a fully open-source project, SIPdb is designed to support and grow with this evolving ecosystem, serving as a scalable platform for meta-analysis, cross-study synthesis, and hypothesis generation in microbial ecology.

Current limitations and methodological evolution

While SIPdb represents a step forward in enabling cross-study synthesis and application of reverse ecology with SIP-derived sequence data, several limitations of its current form should be acknowledged. First, the database is currently restricted to V4 region amplicon data which has limited phylogenetic resolving power. This limits its utility for identifying fine-scale functional variation, such as strain-level population dynamics or accessory gene content, which are increasingly recognized as important drivers of microbial ecology and biogeochemical function [73–75]. Expanding the SIPdb to include shotgun metagenomic and metatranscriptomic datasets will be essential for improving both taxonomic resolution and direct functional inference. However, such data were still relatively scarce at the time of publication. The SQLite database and RShiny dashboard were designed with modularity and extensibility in mind, allowing users to apply different differential abundance frameworks and customize enrichment detection strategies.

One of these frameworks is quantitative SIP (qSIP), which estimates atom-percent excess from per-fraction gene copy numbers [31]. Although the HTS-SIP framework used in our pipeline supports qSIP [50], incorporator designation is sufficient for the cross-study reverse-ecology applications SIPdb targets. Most users need to know which taxa incorporated a given substrate, not the relative rates at which they did so. Quantitative atom-percent excess incorporation rates are most informative for within-experiment comparisons among co-occurring populations, a distinct analytical goal better addressed by dedicated qSIP tools. Consequently, the SIPdb is well appointed to serve the majority of the SIP community, where conventional SIP accounts for roughly 94% of the SIP literature (qSIP ~6%; Fig. S1). The current BLAST-based search architecture may need to evolve toward more scalable alignment and indexing tools, as well as more platforms with greater computational resources like Galaxy [76] or KBase [77].

An additional limitation lies in the abstraction of study-specific methodological detail into simplified categorical metadata fields used in designating the isotopic incorporators that form the basis of the SIPdb. While necessary for interoperability, this standardization introduces a degree of data loss, particularly around incubation conditions, substrate labeling strategies, and gradient fractionation schemes. These factors can influence the identity and detectability of incorporators, potentially biasing results presented in the SIPdb toward taxa exhibiting the strongest incorporation effects. This tradeoff between standardization and preserving useful contextual information reflects a broader challenge in microbiome database development, where ease of use and broad applicability must be balanced against preserving nuanced experimental context. Future iterations of SIPdb could include layered metadata or filterable provenance flags that allow expert users to drill down into methodological differences while retaining an accessible default view for simple exploratory analyses.

Acknowledgements

We thank Kaitlyn Johnson and Payton Taylor, at Purdue University, for their help curating the list of SIP studies.

Supplementary material

Supplementary material is available at *ISME Communications* online.

Conflicts of interest

None declared.

Funding

Funding support for ABT was provided by the Brazilian Ministry of Education and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) Fellowship Program (<https://ror.org/00x0ma614>). Partial funding support for RCW was provided by the USDA National Institute of Food and Agriculture (NIFA; <https://ror.org/03wjh4t84>) Hatch Grant program (IND90024429), administered by Purdue University (<https://ror.org/05p8z3f47>). SJB's effort was supported by the Department of Energy (DOE) Office of Science, Office of Biological and Environmental Research (BER) Early Career Research Project (award SCW1883). JAK's effort was supported by the LLNL "Microbes Persist" Soil Microbiome Science Focus Area, funded by the U.S. DOE Office of Science, BER Genomic Science program (award SCW1632). Work at Lawrence Livermore National Laboratory was performed under U.S. Department of Energy Contract DE-AC52-07NA27344.

Data availability

The SIPdb analysis workflow (phyloseq objects and R scripts), the SQLite database and sequence data, and the dashboard app ("app.R") are all open source and available for local use or adaptation to future use cases via GitHub (<https://github.com/alexbtrentin/sipdb-pipeline>). All additional study metadata are summarized in Table S2. SIPSim simulation information is summarized in Table S6 (reference genome inputs), Table S7 (simulation design parameters), and Table S8 (run settings). Intermediary files used for SIPdb analysis and queries are archived on The Open Science Framework (OSF) under the accession: <https://osf.io/cwg6r/overview>.

References

1. Djemiel C, Maron PA, Terrat S, *et al.* Inferring microbiota functions from taxonomic genes: a review. *GigaScience* 2022;**11**:giab090. <https://doi.org/10.1093/gigascience/giab090>
2. DeLong EF. Microbial population genomics and ecology: the road ahead. *Environ Microbiol* 2004;**6**:875–8. <https://doi.org/10.1111/j.1462-2920.2004.00668.x>
3. Lin JN, Lai CH, Lin SY, *et al.* Effect of intragenomic sequence heterogeneity among multiple 16S rRNA genes on species identification of *Elizabethkingia*. *Microbiol Spectrum* 2022;**10**:e01338–22. <https://doi.org/10.1128/spectrum.01338-22>
4. Hong DND, Long TH, Tien NTT, *et al.* A multi locus sequence analysis scheme for phylogeny of the *Bacillus subtilis* species complex and its advantages over 16S rRNA genes. *HCMCOUJS - ENGINEERING AND TECHNOLOGY* 2022;**12**:22–33. <https://doi.org/10.46223/HCMCOUJS.tech.en.12.2.2255.2022>
5. Regueira-Iglesias A, Balsa-Castro C, Blanco-Pintos T, *et al.* Critical review of 16S rRNA gene sequencing workflow in microbiome studies: from primer selection to advanced data analysis. *Mol Oral Microbiol* 2023;**38**:347–99. <https://doi.org/10.1111/omi.12434>

6. Matchado MS, Rühlemann M, Reitmeier S, *et al.* On the limits of 16S rRNA gene-based metagenome prediction and functional profiling. *Microb Genom* 2024;**10**:001203. <https://doi.org/10.1099/mgen.0.001203>
7. Parks DH, Chuvochina M, Rinke C, *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* 2022;**50**:D785–94. <https://doi.org/10.1093/nar/gkab776>
8. Levy R, Borenstein E. Reverse ecology: from systems to environments and back. In: Soyer OS (ed.), *Evolutionary Systems Biology. Adv Exp Med Biol.* New York, NY: Springer New York 2012; **751**:329–45. https://doi.org/10.1007/978-1-4614-3567-9_15
9. Arevalo P, VanInsberghe D, Elsherbini J, *et al.* A reverse ecology approach based on a biological definition of microbial populations. *Cell* 2019;**178**:820–834.e14. <https://doi.org/10.1016/j.cell.2019.06.033>
10. Riddle M, Hepp S, Hardeep F, *et al.* Differential roles of deterministic and stochastic processes in structuring soil bacterial ecotypes across terrestrial ecosystems. *Nat Commun* 2025;**16**:2337. <https://doi.org/10.1038/s41467-025-57526-x>
11. Jia J, De Goede R, Li Y, *et al.* Unlocking soil health: are microbial functional genes effective indicators? *Soil Biol Biochem* 2025;**204**:109768. <https://doi.org/10.1016/j.soilbio.2025.109768>
12. Wilhelm RC, Amsili JP, Kurtz KSM, *et al.* Ecological insights into soil health according to the genomic traits and environment-wide associations of bacteria in agricultural soils. *ISME Commun* 2023;**3**:1. <https://doi.org/10.1038/s43705-022-00209-1>
13. Deel HL, Manter DK, Moore JM. Rethreading the needle: a novel molecular index of soil health (MISH) using microbial functional genes to predict soil health management. *PLoS One* 2024;**19**:e0314072. <https://doi.org/10.1371/journal.pone.0314072>
14. Lassalle F, Muller D, Nesme X. Ecological speciation in bacteria: reverse ecology approaches reveal the adaptive part of bacterial cladogenesis. *Res Microbiol* 2015;**166**:729–41. <https://doi.org/10.1016/j.resmic.2015.06.008>
15. Lagkouravdos I, Joseph D, Kapfhammer M, *et al.* IMNGS: a comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. *Sci Rep* 2016;**6**:33721. <https://doi.org/10.1038/srep33721>
16. Rocca JD, Simonin M, Blaszcak JR, *et al.* The microbiome stress project: toward a global meta-analysis of environmental stressors and their effects on microbial communities. *Front Microbiol* 2019;**9**:3272. <https://doi.org/10.3389/fmicb.2018.03272>
17. Ramirez KS, Knight CG, De Hollander M, *et al.* Detecting macroecological patterns in bacterial communities across independent studies of global soils. *Nat Microbiol* 2017;**3**:189–96. <https://doi.org/10.1038/s41564-017-0062-x>
18. Thompson LR, Sanders JG, McDonald D, *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 2017;**551**:457–63. <https://doi.org/10.1038/nature24621>
19. Neufeld JD, Wagner M, Murrell JC. Who eats what, where and when? Isotope-labelling experiments are coming of age. *ISME J* 2007;**1**:103–10. <https://doi.org/10.1038/ismej.2007.30>
20. Dumont MG, Hernández García M (eds). *Stable Isotope Probing: Methods and Protocols.* New York, NY: Springer New York, 2019; Vol. **2046** <https://doi.org/10.1007/978-1-4939-9721-3>.
21. Gülay A, Fowler SJ, Tatari K, *et al.* DNA- and RNA-SIP reveal *Nitrospira* spp. as key drivers of nitrification in groundwater-fed biofilters. *mBio* 2019;**10**:e01870–19. <https://doi.org/10.1128/mBio.01870-19>
22. Nieto EE, Jurburg SD, Steinbach N, *et al.* DNA stable isotope probing reveals the impact of trophic interactions on bioaugmentation of soils with different pollution histories. *Microbiome* 2024;**12**:146. <https://doi.org/10.1186/s40168-024-01865-2>
23. Lemmel F, Maunoury-Danger F, Leyval C, *et al.* DNA stable isotope probing reveals contrasted activity and phenanthrene-degrading bacteria identity in a gradient of anthropized soils. *FEMS Microbiol Ecol* 2019;**95**:fiz181. <https://doi.org/10.1093/femsec/fiz181>
24. Hernández M, Roy S, Keevil CW, *et al.* Identification of diverse antibiotic resistant bacteria in agricultural soil with H218O stable isotope probing combined with high-throughput sequencing. *Environ Microbiome* 2023;**18**:34. <https://doi.org/10.1186/s40793-023-00489-7>
25. Liu R, Han C, Kang Y, *et al.* Active microbial taxa preferentially assimilate inorganic nitrogen in acidic upland soils using a 15N-DNA-SIP approach. *Soil Biol Biochem* 2024;**188**:109226. <https://doi.org/10.1016/j.soilbio.2023.109226>
26. Sun H, Huang K, Zhang X, *et al.* Stable isotope probing reveals specific assimilating bacteria of refractory organic compounds in activated sludge. *Water Res* 2022;**212**:118105. <https://doi.org/10.1016/j.watres.2022.118105>
27. Pepe-Ranney C, Campbell AN, Koechli CN, *et al.* Unearthing the ecology of soil microorganisms using a high resolution DNA-SIP approach to explore cellulose and xylose metabolism in soil. *Front Microbiol* 2016;**7**:703. <https://doi.org/10.3389/fmicb.2016.00703>
28. Foley MM, Stone BWG, Caro TA, *et al.* Growth rate as a link between microbial diversity and soil biogeochemistry. *Nat Ecol Evol* 2024;**8**:2018–26. <https://doi.org/10.1038/s41559-024-02520-7>
29. Hungate BA, Marks JC, Power ME, *et al.* The functional significance of bacterial predators. *mBio* 2021;**12**:e00466–21. <https://doi.org/10.1128/mBio.00466-21>
30. Simpson A, Wood-Charlson EM, Smith M, *et al.* MISIP: a data standard for the reuse and reproducibility of any stable isotope probing-derived nucleic acid sequence and experiment. *GigaScience* 2024;**13**:giae071. <https://doi.org/10.1093/gigascience/giae071>
31. Hungate BA, Mau RL, Schwartz E, *et al.* Quantitative microbial ecology through stable isotope probing. *Appl Environ Microbiol* 2015;**81**:7570–81. <https://doi.org/10.1128/AEM.02280-15>
32. Caporaso JG, Lauber CL, Walters WA, *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A* 2011;**108**:4516–22. <https://doi.org/10.1073/pnas.100080107>
33. Bolyen E, Rideout JR, Dillon MR, *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 2019;**37**:852–7. <https://doi.org/10.1038/s41587-019-0209-9>
34. Quast C, Pruesse E, Yilmaz P, *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2012;**41**:D590–6. <https://doi.org/10.1093/nar/gks1219>
35. R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria; 2025. <https://www.R-project.org/>
36. McMurdie PJ, Holmes S. Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 2013;**8**:e61217. <https://doi.org/10.1371/journal.pone.0061217>
37. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550. <https://doi.org/10.1186/s13059-014-0550-8>
38. Chen Y, Chen L, Lun ATL, *et al.* edgeR v4: powerful differential analysis of sequencing data with expanded functionality and improved

- support for small counts and larger datasets. *Nucleic Acids Res* 2025;**53**:gkaf018. <https://doi.org/10.1093/nar/gkaf018>
39. Ritchie ME, Phipson B, Wu D, *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47–7. <https://doi.org/10.1093/nar/gkv007>
 40. Fernandes AD, Macklaim JM, Linn TG, *et al.* ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. *PLoS One* 2013;**8**:e67019. <https://doi.org/10.1371/journal.pone.0067019>
 41. Youngblut ND, Buckley DH. Intra-genomic variation in G + C content and its implications for DNA stable isotope probing. *Environ Microbiol Rep* 2014;**6**:767–75. <https://doi.org/10.1111/1758-2229.12201>
 42. Youngblut ND, Barnett SE, Buckley DH. SIPSim: a Modeling toolkit to predict accuracy and aid design of DNA-SIP experiments. *Front Microbiol* 2018;**9**:570. <https://doi.org/10.3389/fmicb.2018.00570>
 43. Chang W, Cheng J, Allaire J. *et al.* Shiny: Web Application Framework for R. Vienna, Austria: Comprehensive R Archive Network (CRAN) 2012; v1.13.0. <https://doi.org/10.32614/CRAN.package.shiny>
 44. Wang M, Chen X, Tang Y, *et al.* Substrate availability and toxicity shape the structure of microbial communities engaged in metabolic division of labor. *mLife* 2022;**1**:131–45. <https://doi.org/10.1002/mlf2.12025>
 45. Li Z, Eshghdoostkhatami Z, Cupples AM. Biodegradation rates of low concentrations of 1,4-dioxane in impacted site sediments and agricultural soils and an investigation of the microorganisms and genes involved. *Environ Pollut* 2025;**381**:126513. <https://doi.org/10.1016/j.envpol.2025.126513>
 46. Garritano AN, Song W, Thomas T. Carbon fixation pathways across the bacterial and archaeal tree of life. *PNAS Nexus* 2022;**1**:pgac226. <https://doi.org/10.1093/pnasnexus/pgac226>
 47. Stultiens K, Van Kessel MAHJ, Frank J, *et al.* Diversity, enrichment, and genomic potential of anaerobic methane- and ammonium-oxidizing microorganisms from a brewery wastewater treatment plant. *Appl Microbiol Biotechnol* 2020;**104**:7201–12. <https://doi.org/10.1007/s00253-020-10748-z>
 48. Jeckelmann JM, Erni B. Transporters of glucose and other carbohydrates in bacteria. *Pflugers Arch - Eur J Physiol* 2020;**472**:1129–53. <https://doi.org/10.1007/s00424-020-02379-0>
 49. Barnett SE, Youngblut ND, Buckley DH. Bacterial community dynamics explain carbon mineralization and assimilation in soils of different land-use history. *Environ Microbiol* 2022;**24**:5230–47. <https://doi.org/10.1111/1462-2920.16146>
 50. Youngblut ND, Barnett SE, Buckley DH. HTSSIP: an R package for analysis of high throughput sequencing data from nucleic acid stable isotope probing (SIP) experiments. *PLoS One* 2018;**13**:e0189616. <https://doi.org/10.1371/journal.pone.0189616>
 51. Yang L, Chen J. Benchmarking differential abundance analysis methods for correlated microbiome sequencing data. *Brief Bioinform* 2023;**24**:bbac607. <https://doi.org/10.1093/bib/bbac607>
 52. McNichol SM, Sanchez-Quete F, Loeb SK, *et al.* Dynamics of carbon substrate competition among heterotrophic microorganisms. *ISME J* 2024;**18**:wrae018. <https://doi.org/10.1093/ismejo/wrae018>
 53. Alothman A, Duarte CM, Qurban MA, *et al.* Flow of heterotrophic production in oligotrophic ocean waters. *Front Microbiol* 2025;**16**:1530627. <https://doi.org/10.3389/fmicb.2025.1530627>
 54. Thakur SS, González SS, Suravajhala P, *et al.* Metagenomic insights of microbial functions under conventional and conservation agriculture. *World J Microbiol Biotechnol* 2025;**41**:100. <https://doi.org/10.1007/s11274-025-04312-y>
 55. Osborne CD, Haritos VS. Beneath the surface: evolution of methane activity in the bacterial multicomponent monooxygenases. *Mol Phylogenet Evol* 2019;**139**:106527. <https://doi.org/10.1016/j.ympev.2019.106527>
 56. Nuccio EE, Blazewicz SJ, Lafler M, *et al.* HT-SIP: a semi-automated stable isotope probing pipeline identifies cross-kingdom interactions in the hyphosphere of arbuscular mycorrhizal fungi. *Microbiome* 2022;**10**:199. <https://doi.org/10.1186/s40168-022-01391-z>
 57. Mooshammer M, Kitzinger K, Schintlmeister A, *et al.* Flow-through stable isotope probing (flow-SIP) minimizes cross-feeding in complex microbial communities. *ISME J* 2021;**15**:348–53. <https://doi.org/10.1038/s41396-020-00761-5>
 58. DeRito CM, Pumphrey GM, Madsen EL. Use of field-based stable isotope probing to identify adapted populations and track carbon flow through a phenol-degrading soil microbial community. *Appl Environ Microbiol* 2005;**71**:7858–65. <https://doi.org/10.1128/AEM.71.12.7858-7865.2005>
 59. Uhlik O, Leewis MC, Strejcek M, *et al.* Stable isotope probing in the metagenomics era: a bridge towards improved bioremediation. *Biotechnol Adv* 2013;**31**:154–65. <https://doi.org/10.1016/j.biotechadv.2012.09.003>
 60. Martiny AC, Treseder K, Pusch G. Phylogenetic conservatism of functional traits in microorganisms. *ISME J* 2013;**7**:830–8. <https://doi.org/10.1038/ismej.2012.160>
 61. Viladomat Jasso M, García-Ulloa M, Zapata-Peñasco I, *et al.* Metagenomic insight into taxonomic composition, environmental filtering and functional redundancy for shaping worldwide modern non-lithifying microbial mats. *PeerJ* 2024;**12**:e17412. <https://doi.org/10.7717/peerj.17412>
 62. Mo Y, Zhang W, Yang J, *et al.* Biogeographic patterns of abundant and rare bacterioplankton in three subtropical bays resulting from selective and neutral processes. *ISME J* 2018;**12**:2198–210. <https://doi.org/10.1038/s41396-018-0153-6>
 63. Li YQ, Xin Y, Li C, *et al.* Metagenomics-metabolomics analysis of microbial function and metabolism in petroleum-contaminated soil. *Braz J Microbiol* 2023;**54**:935–47. <https://doi.org/10.1007/s42770-023-01000-7>
 64. Hayasaka M, Hamajima L, Yoshida Y, *et al.* Phenanthrene degradation by a flavoprotein monooxygenase from *Phanerochaete chrysosporium*. *Appl Environ Microbiol* 2025;**91**:e01574–24. <https://doi.org/10.1128/aem.01574-24>
 65. Silva NM, Romagnoli CL, Santiago CRDN, *et al.* Multi-approach characterization of novel pyrene-degrading *Mycolicibacterium austroafricanum* isolates lacking nid genes. *Microorganisms* 2023;**11**:1413. <https://doi.org/10.3390/microorganisms11061413>
 66. Mahendra S, Alvarez-Cohen L. Kinetics of 1,4-dioxane biodegradation by monooxygenase-expressing bacteria. *Environ Sci Technol* 2006;**40**:5435–42. <https://doi.org/10.1021/es060714v>
 67. Deng D, Li F, Li M. A novel propane monooxygenase initiating degradation of 1,4-dioxane by *Mycobacterium dioxanotrophicus* PH-06. *Environ Sci Technol Lett* 2018;**5**:86–91. <https://doi.org/10.1021/acs.estlett.7b00504>
 68. Masuda H, McClay K, Steffan RJ, *et al.* Biodegradation of tetrahydrofuran and 1,4-dioxane by soluble diiron monooxygenase in *Pseudonocardia* sp. Strain ENV478. *Microb Physiol* 2012;**22**:312–6. <https://doi.org/10.1159/000343817>
 69. Lemos LN, De Carvalho FM, Gerber A, *et al.* Genome-centric metagenomics reveals insights into the evolution and metabolism of a new free-living group in Rhizobiales. *BMC Microbiol* 2021;**21**:294. <https://doi.org/10.1186/s12866-021-02354-4>
 70. Teng Y, Wang X, Li L, *et al.* Rhizobia and their bio-partners as novel drivers for functional remediation in contaminated soils. *Front Plant Sci* 2015;**6**:32. <https://doi.org/10.3389/fpls.2015.00032>
 71. Soo RM, Woodcroft BJ, Parks DH, *et al.* Back from the dead; the curious tale of the predatory cyanobacterium *Vampirovibrio chlorellavorus*. *PeerJ* 2015;**3**:e968. <https://doi.org/10.7717/peerj.968>

72. Vyshenska D, Sampara P, Singh K, *et al.* A standardized quantitative analysis strategy for stable isotope probing metagenomics. *mSystems* 2023;**8**:e01280-22. <https://doi.org/10.1128/msystems.01280-22>
73. Thomas F, Corre E, Cébron A. Stable isotope probing and metagenomics highlight the effect of plants on uncultured phenanthrene-degrading bacterial consortium in polluted soil. *ISME J* 2019;**13**:1814–30. <https://doi.org/10.1038/s41396-019-0394-z>
74. Wilhelm RC, Pepe-Ranney C, Weisenhorn P, *et al.* Competitive exclusion and metabolic dependency among microorganisms structure the cellulose economy of an agricultural soil. *mBio* 2021;**12**:e03099–20. <https://doi.org/10.1128/mBio.03099-20>
75. Ziels RM, Sousa DZ, Stensel HD, *et al.* DNA-SIP based genome-centric metagenomics identifies key long-chain fatty acid-degrading populations in anaerobic digesters with different feeding frequencies. *ISME J* 2018;**12**:112–23. <https://doi.org/10.1038/ismej.2017.143>
76. The Galaxy Community, Abueg LAL, Afgan E, *et al.* The galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Res* 2024;**52**:W83–94. <https://doi.org/10.1093/nar/gkae410>
77. Arkin AP, Cottingham RW, Henry CS, *et al.* KBase: the United States Department of Energy Systems Biology Knowledgebase. *Nat Biotechnol* 2018;**36**:566–9. <https://doi.org/10.1038/nbt.4163>