Research Square

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

Eukaryotic mosaics across sediment, sand and soil: local diversity and habitat associations through 18S profiling

Minji Kim
Jeju National University
So-Jeong Kim
Korea Institute of Geoscience and Mineral Resources
Soo-Je Park
sjpark@jejunu.ac.kr

Jeju National University

Research Article

Keywords: 18S rRNA gene, sediment, sand, soil, eukaryotic community, differential abundance, cross-study classification

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11521166/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Sediment, sand and soil harbor heterogeneous eukaryotic assemblages. An 18S ribosomal RNA gene survey characterizes 15 Korean solid-substrate samples and compares them with 125 external libraries.

Results

At 20,000 reads per local sample, richness ranges from 73 to 1,034 amplicon sequence variants (ASVs), with contrasting animal, fungal and algal contributions. Phototroph-associated reads comprise 0.56–93.10% of the protist-enriched fraction. Sixty-six abundance-selected ASVs contribute 57.02% of pooled reads, yet no ASV in the complete table occurs in all 15 samples. Air-exposed versus permanently submerged grouping accounts for 5.86% of local compositional variation (P = 0.9335). Across 140 libraries, dataset-adjusted sample-type variation is 9.77–12.04% at Genus and Family levels under within-dataset permutations; finer blocking retains an association alongside dispersion differences. Source-cluster-adjusted differential abundance models identify exploratory between-study associations for Oxytrichidae and Sandonidae, with lower model-estimated abundance in marine water than land sediment; no source study contains both habitats. Whole-study-held-out classification transfers unevenly, with three-class balanced accuracy of 0.479–0.486 and frequent assignment of soil and water to sediment/sand.

Conclusions

The inventory distinguishes local dominance from recurrence. External comparisons nominate lineage–sample-type associations for independent testing, without establishing source-independent habitat effects or validated ecological indicators.

1. Background

Sediment, sand and soil harbour eukaryotic assemblages with different contributions from animals, fungi, algae and other protists. Their distributions can guide studies of primary production, consumption, decomposition and host associations. Comparisons across soil, freshwater and marine ecosystems have revealed differences in protist composition and taxonomy-inferred trophic profiles [1]. Soil surveys have further related contrasts among consumer-, phototroph- and parasite-associated lineages to potential differences in energy acquisition and host relationships [2]. We therefore distinguish dominant contributors within samples from community features that recur across samples and environmental contexts.

Heterogeneous local assemblages may coexist with recurring sample-type associations, but those associations need not predict sample context in a different study. Oliverio et al. [2] linked soil protist composition most strongly to precipitation and identified geographic and biome sampling gaps. Their findings motivate connecting local inventories to broader environmental comparisons, without assuming that relationships found in soils also apply to sediment and sand. The present survey first characterizes variation among local assemblages, then asks whether sample-type associations extend across the assembled references and remain informative when an entire source study is withheld.

We compared air-exposed and permanently submerged materials. Soil protist surveys have associated community variation with climatic moisture and edaphic properties, with different responses among lineages [2, 3]. Lake surveys have linked protist assemblages to trophic state and watershed context [4]. Exposure does not directly measure moisture, oxygen or light, and samples within either exposure category may differ substantially in composition. Examining individual-sample diversity and dominant lineages alongside the recorded exposure contrast can therefore reveal variation that a single group label does not summarize.

Sequence sharing addresses a complementary question. Pooled abundance describes how much a sequence contributes to recovered reads, whereas recurrence describes how often it is detected across samples. A sequence may attain high pooled abundance through a large contribution in one sample without being widely shared. Conversely, recurrence of a higher taxon need not mean that its constituent sequences are shared. Limited sequence overlap between soil and freshwater protist inventories illustrates the importance of distinguishing these levels of comparison [5]. Amplicon sequence variants (ASVs) offer reproducible sequence-defined units for examining abundance and recurrence within a consistently processed collection [6, 7].

Broad 18S ribosomal RNA (rRNA) gene profiling connects these questions across major eukaryotic lineages. The Protist Ribosomal Reference database (PR2) and environmental sequence resources support comparison at explicit taxonomic ranks [8, 9]. Retaining animal, fungal, plant and protist sequences describes the recovered eukaryotic assemblage, while a protist-enriched view can resolve contributions that are less apparent in whole-library summaries. Interpreting these views together distinguishes broad heterotrophic carbon affinity from particular modes of feeding or host association. The interpretation remains sequence-based: rDNA copy number varies among eukaryotes, and primer region and sequence length affect recovered profiles and taxonomic resolution [10, 11, 12, 13]. Cross-study comparisons consequently require compatible feature definitions rather than equating independently assigned ASV labels or interpreting read fractions as cell fractions.

External datasets allow us to assess whether local community features also distinguish environmental contexts beyond the sampled locations. Whole-community comparisons can identify sample-type-associated variation, while differential abundance analysis (DAA) can identify lineages whose sequence representation differs between specified groups. Such lineages are candidates for independent ecological testing, not established indicators. We used Random Forest (RF) classification to test whether

community profiles could predict the sample type of material from a study withheld during training. A model that distinguishes samples within the assembled data need not do so for a new source study, making validation design part of the ecological interpretation [14]. External references broaden the environmental comparison; they do not supply field replication at the original locations.

We used this two-stage approach to characterize 15 Korean sediment, sand and soil samples, termed KSS-18S, and compare them with 75 libraries from a Coastal reference [15] and 50 selected EukBank V4 libraries [16]. Locally, we asked how diversity, dominant lineages and potential trophic affinities varied among samples, whether diversity and composition differed between exposure categories, and whether abundant ASVs were also recurrent. Across datasets, we asked whether broader sample-type structure was accompanied by specific taxonomic associations and whether community profiles could classify materials from withheld source studies. Alpha and beta diversity, taxonomic composition and ASV distributions addressed the local questions, with protist-enriched composition, potential trophic affinities and distribution breadth providing descriptive context. This local inventory is the primary contribution. Genus- and Family-level comparisons evaluated exploratory taxonomic associations across datasets and prediction of sample types from withheld studies.

2. Methods

2.1 Local sampling and the role of the comparative datasets

KSS-18S comprised 15 independently collected samples from distinct locations, each represented by one library. The survey profiled DNA recovered from bulk sediment, sand and soil; solid-substrate-associated refers to these sampled materials, not a selectively isolated attached-cell fraction. Each location was sampled once, without within-location field replicates. Twelve locations were on Jeju Island and three on the Korean Peninsula; collection campaigns occurred in spring or autumn between 2019 and 2023. The six air-exposed materials comprised sediments DBW, SSR and SWP and soils GJS, HFS and HMC. The nine permanently submerged materials comprised sediments DGW, GKP, HFP, HGP, NSP, OGP and SSP and sands GC and HDP. Permanent submergence refers to submerged sediment layers, rather than temporary water cover. Substrate and exposure were partly aligned because every soil was exposed and both sands were submerged. Region denoted the geographic division between Jeju Island and the Korean Peninsula (Table S1; Figure S1).

For KSS samples, pH was measured in a 1:5 (w/v) soil-to-distilled-water suspension. The soil-pH method reference [17] and the scope of the recorded procedure are specified in SI Materials and Methods S1. The analysis metadata contained five numerical point records, ten range records and the supplied categories acid, weak acid and neutral. The full collection contained seven, five and three samples in these categories; the submerged subset contained three, four and two. Tests used the categories, while numerical point and range-derived values supported separate descriptive profiles (Tables S1–S2; SI Materials and Methods S1 and S3). Matched moisture, redox, organic matter, grain-size and temperature

records were unavailable. A water-column library, SWPW, supplied context in selected displays but was not included in any 15-sample group analysis.

The Coastal dataset supplies a replicated reference comparison of sand/sediment and water under a common study design [15]. It contains 25 site–season–sample-type combinations from seven sites, comprising 13 sand/sediment and 12 water combinations. Three independently collected spatial biological replicates per combination yield 75 libraries: 39 sand/sediment and 36 water. We retained these libraries as biological observations rather than collapsing them to 25 averaged samples. The published environmental measurements describe spring sea-sand at seven sites, not individual water or autumn samples. The existing pH summary links these records to site-average compositions and is interpreted only as site context. The present comparison uses existing DADA2-reprocessed library tables; the effect estimates reported below come from this comparative analysis, not from transcription of the published results. Under the published protocol, water represented a nominal 0.22–20-μm suspended fraction, whereas sea-sand was processed without the same size fractionation.

The EukBank comparison used the environmental 18S V4 resource of Berney et al. [16] and was restricted to 50 selected libraries from 32 source studies. It included ten libraries in each of five operational categories: land sediment, land soil, freshwater, marine sediment and marine water. Selection required the predefined V4-primer annotation, sufficient reads and geographic coordinates; a deterministic geographic-spread procedure limited source-study representation within categories. The analysis used this subset, not the complete collection. The exact EukBank snapshot and retrieval date were not established in the available records. No selected library had a pH record, and salinity was recorded for 12 of the 50. The selection sequence and limits are documented in SI Materials and Methods S1 and Additional Data Table AD1.

The comparative design thus comprised 140 libraries in three datasets: 15 KSS-18S, 75 Coastal and 50 EukBank. KSS soil was mapped to land soil, other KSS solids to land sediment, and Coastal sea sand to marine sediment. These operational categories identify sample context, not measured salinity or an isolated attached-cell fraction. Dataset membership, source-study identifiers, category mapping and assignment coverage are retained in Table S3. The external collections broaden the comparison without providing field replicates of KSS locations.

2.2 DNA preparation and 18S sequencing

For KSS-18S, DNA was extracted with the DNeasy PowerSoil Kit (QIAGEN, Hilden, Germany). We amplified the 18S rRNA gene V4 region by polymerase chain reaction (PCR) with Herculase II Fusion DNA Polymerase, using primers V4F (5'-CCAGCASCYGCGGTAATTCC-3') and V4R (5'-ACTTTCGTTCTTGAT-3'). Triplicate PCR products were pooled for each sample; these were technical amplification replicates, not additional biological samples. Macrogen (Seoul) prepared and sequenced indexed libraries on an Illumina MiSeq using paired-end 2 × 300-bp reads. The PCR-negative control produced no visible amplification band on a gel. Negative controls were not sequenced. SI Materials and

Methods S1 records the laboratory procedure; Coastal methods are also described in its published study [15].

2.3 Sequence inference, taxonomy and the KSS common-depth table

Cutadapt v5.2 removed primers by anchored 5' matching with a maximum error rate of 0.10, minimum overlap of 15 nt and no indels. Paired-mode processing required both expected primers and retained reads at least 100 nt long. DADA2 1.38.0 filtering used truncLen 250/220, maxEE 2/4, maxN 0 and truncQ 2 [6]. Forward and reverse errors were learned separately, and sequence inference was independent by sample. Pairs were merged with at least 20 nt overlap and zero mismatches; consensus chimera removal followed. Taxonomy was assigned against PR2 v5.1.1 with minimum bootstrap support 50 [8]. All PR2-assigned Eukaryota, including Metazoa and nuclear Chloroplastida, were retained. SI Materials and Methods S2 specifies the reported processing environments. KSS read retention is summarized in Figure S1; common-depth alpha diversity and Good's coverage are reported in Table S4. Additional Data Table AD1 distinguishes documented methods, recovered selection records and unresolved upstream provenance.

For KSS diversity, pooled abundance and recurrence, each library was subsampled once without replacement to 20,000 reads using a fixed random-number seed recorded in SI Materials and Methods S2. This produced the 300,000-read common-depth table. The table supplies the local sequence distributions and is distinct from the repeated subsampling used for comparative distance sensitivity below. Rarefaction and common-depth coverage are documented in Figure S2 and Table S4; the local standardization does not imply complete recovery of organismal richness.

2.4 Diversity and community comparisons within KSS

We described alpha diversity with observed ASVs, Shannon diversity using natural logarithms, Simpson diversity ($1 - \Sigma p^2$) and Pielou's evenness. Evenness was not separately hypothesis-tested. Good's coverage, calculated from singleton counts, summarizes sequence coverage rather than completeness of the organismal inventory. Exposure and submerged-pH comparisons used exact Kruskal–Wallis tests over all 5,005 and 1,260 distinct label allocations, respectively. Epsilon-squared quantified effect size. We applied Benjamini–Hochberg (BH) adjustment within each recorded test family [18]. The six tested estimators were observed ASVs, Shannon, Simpson, inverse Simpson, Chao1 and ACE. Substrate comparisons and air-exposed pH summaries remained descriptive because of sparse or confounded category combinations (SI Materials and Methods S3; Tables S5–S6).

Bray–Curtis dissimilarities described beta diversity and were displayed by principal coordinates analysis (PCoA) [19]. Exposure and submerged-pH permutational multivariate analysis of variance (PERMANOVA) used their respective exact allocation sets; dispersion diagnostics used 9,999 permutations [20, 21]. The contextual ordination in Fig. 1D contains 16 libraries, including SWPW, whereas the KSS group tests use only the 15 solid-substrate libraries. A separate exploratory Genus- and

Family-level PERMANOVA compared the three soils with the 12 other solids using unrestricted permutations. That contrast is distinguished from exposure and from comparative DAA categories in Tables S7–S8; its comparative-summary BH correction is specified in SI Materials and Methods S4.

2.5 Sequence sharing and supporting taxonomic profiles

We defined ASV occurrence as at least one read in a common-depth library and recurrence as the number of libraries meeting that criterion. The high-recurrence summary required at least eight of 15 samples; ubiquity required all 15. Dominance was defined independently: an ASV had to contribute at least 0.2% of the pooled 300,000 reads, equivalent to 600 reads. The abundance-selected ASVs were aligned using MAFFT v7.526 and displayed as an unrooted FastTree v2.2.0 V4 sequence tree [22, 23]. This display provides sequence-similarity context for selected ASVs, not an organismal phylogeny of the whole community (SI Materials and Methods S2; Tables S9–S10).

To describe the protist-enriched component, we excluded Metazoa, Fungi, Embryophyceae and reads with insufficiently resolved Division or Opisthokonta context. Non-land-plant algae remained, so the subset was not defined as exclusively unicellular. We described lifestyle-related composition as potential trophic affinities assigned by conservative lineage-specific rules, with heterotroph-associated, phototroph-associated, narrowly assigned mixotroph-associated, variable/insufficient-resolution and unresolved-taxonomy categories [24]. For animal and fungal reads, heterotroph-associated denotes carbon affinity rather than a specific feeding guild. Taxonomic assignment and trophic resolution were recorded separately; a Class-level identification did not necessarily support a single trophic assignment (Table S11).

Whole-library profiles used all 20,000 reads per sample as the denominator, whereas within-subset Class and trophic profiles used only retained protist-enriched reads. Pooled summaries divided category read totals by 300,000 whole-eukaryote reads or 113,530 retained protist-enriched reads, respectively. The latter is a read-weighted pooled summary, not an equal-sample mean of within-subset percentages. Apicomplexa, Perkinsida, Phytomyxea and Pirsoniales were tracked separately as selected parasite- or endobiont-associated lineages; their read fractions were not added to the mutually exclusive trophic categories. We did not perform trophic-group hypothesis tests or protist-only diversity, DAA or RF analyses. PR2 rank names, including Class, Family and Genus, are used consistently as database ranks (Figures S3–S4; Table S11).

Standardized Levins breadth summarized the evenness of read distributions across observed samples or categories [25]. The sample-based calculation retained 94 Class-level features detected in at least two KSS samples, including an unresolved-Class feature. Category summaries instead retained 93 assigned Classes and gave categories equal weight. Class-abundance-weighted pH means and standard deviations summarized the eligible records, not physiological optima or tolerance. KSS profiles used its sample records; Coastal profiles linked compositions averaged across all libraries at each site to spring sea-sand pH and therefore describe site context, not sample-matched pH responses. These descriptors

distinguish distribution evenness from detection frequency and do not designate generalists or specialists. Eligibility, denominators and formulas are provided in SI Materials and Methods S3 and Tables S2 and S12.

2.6 Community structure across datasets

Comparative features were Genus or Family assignments qualified by their full PR2 lineage and target-rank bootstrap support of at least 50. Local ASV labels were not used as cross-study identifiers. We excluded explicitly unassigned target-rank features, retained identifiable PR2 placeholders, and normalized composition within assigned reads. Bray–Curtis PCoA used the Lingoes correction; ordinary centred log-ratio principal component analysis (CLR-PCA) used a 10^{-6} addition to relative compositions before transformation [26, 27]. The heatmaps show the 25 features at each rank with the highest mean assigned relative abundance across the 140 libraries. This ranking was separate from the KSS dominant-ASV selection.

The pooled PERMANOVA tested marginal terms in a model containing dataset and sample type, with 9,999 permutations restricted within dataset. Because dataset membership was not exchangeable under that restriction, we did not interpret the dataset-term P value. This restriction does not fully represent the finer nesting of Coastal sites or EukBank studies. Within-reference checks therefore used site and season adjustment with permutations within site-by-season groups for Coastal, and study-restricted permutations for EukBank. The 18 admissible EukBank habitat-label allocations were also evaluated exactly. Table S8 specifies the comparison units and permutation restrictions for each analysis.

An additional sensitivity analysis used the same Genus and Family distance matrices with finer exchangeability blocks: one block for the 15 independently collected KSS locations, Coastal site-by-season blocks, and EukBank source-study blocks. We tested the estimable sample-type component after conditioning on these blocks with partial distance-based redundancy analysis, using 9,999 shared restricted permutations; the fixed seed and allocation matrix are recorded in SI Materials and Methods S4 and Table S8. We also evaluated the original dataset-plus-sample-type formula under the same restriction and applied a median-based, bias-adjusted dispersion diagnostic with a Lingoes correction. Holm adjustment covered the two rank-level P values within each sensitivity-test family. This analysis addressed only the contrasts represented within blocks, not all five sample types independently; the design and complete results are in Table S8 and SI Materials and Methods S4.

To assess potential effects of uneven sequencing effort on community distances [28], we subsampled each of the 140 libraries without replacement 100 times at each of two depths: 14,000 and 10,000 reads per library. Both depths retained all libraries; the minimum available count was 14,997 Eukaryota reads. For each draw, ASV counts were aggregated with unchanged Genus/Family assignment rules and converted to distances; the 100 distance matrices at each depth were then averaged. Repetition reduced dependence on an individual random draw. The fixed 100-draw exercise was not selected by a convergence criterion and does not establish that 100 draws are sufficient for all downstream analyses. These draws were computational repetitions, not extra biological libraries. These distance-based

analyses used R v4.6.1, vegan v2.7-6 and permute 0.9–10. SI Materials and Methods S7 specifies the draw seeds, matched restrictions and dispersion checks (Table S13). This distance-based sensitivity was distinct from the single-draw KSS diversity table and from the unrarefied DAA models.

2.7 Taxon-associated contrasts

Compositional DAA evaluated taxon-associated contrasts separately from whole-community structure. Analysis of compositions of microbiomes with bias correction 2 (ANCOM-BC2) used unrarefied counts to estimate bias-corrected log-abundance contrasts under its sampling-fraction and taxon-bias assumptions [29]. Models without random intercepts contained sample type for KSS and EukBank, sample type plus season for Coastal, and sample type plus dataset for the pooled 140 libraries. Land sediment was the pooled reference. Corresponding clustered models added random intercepts for Coastal site, EukBank study or the combined source-specific cluster. In that combined identifier, each KSS library was its own unit. The KSS ASV model tested exposure rather than the Genus/Family sample-type contrast. Prevalence thresholds and all executed formulas are given in SI Materials and Methods S6.

Reported terms are taxon-by-contrast coefficients, and more than one term can refer to a taxon. BH adjustment used $\alpha = 0.05$ within the recorded model or contrast. In the outputs, a “nominal” term has BH $q < 0.05$ before additional pseudo-count sensitivity screening; the word does not denote an unadjusted P value. This nominal set includes robust terms; nominal-only denotes the BH-significant subset failing the additional screen. A finite term passing both BH adjustment and that screen is termed pseudo-count-robust. We also assessed coefficient estimability, agreement between models with and without random intercepts, and random-effect singularity. Rank-dropped terms were ineligible, and structural-zero flags were recorded separately from finite coefficients. Neither represents directly calibrated cell or biomass abundance (Tables S14–S15).

The ANOVA-like differential expression tool ALDEx2 supplied 25 within-dataset contrast-by-rank comparisons: two Coastal, 20 EukBank and three KSS analyses [30]. The complete contrasts, group sizes and feature filters are listed in SI Materials and Methods S6 and Table S14. These sensitivity comparisons did not include the pooled 140-library design or site/source-study random effects. Each used 128 Monte Carlo instances, the all-feature CLR denominator and expected Welch-test BH values. Between-group CLR differences and standardized effects were distinguished from one another and from ANCOM-BC2 coefficients. The comparative environment was R v4.6.1 with ANCOMBC v2.14.0 and ALDEx2 v1.44.0. The separate version and settings of the KSS Class-level exposure/pH analysis are specified in SI Materials and Methods S6.

2.8 Classification and the scope of source withholding

The primary RF task classified 71 sediment/sand, 13 soil and 56 water libraries from their Genus or Family profiles [31]. Sediment/sand comprised 12 KSS non-soil libraries, 39 Coastal sand/sediment libraries and 20 EukBank land- or marine-sediment libraries. Soil comprised three KSS and ten EukBank libraries; water comprised 36 Coastal and 20 EukBank libraries. We evaluated this same target under two

blocking designs. Leave-one-project-out (LOPO) withheld one entire source study at a time across 34 studies: KSS, Coastal and 32 EukBank studies. Here “project” means source study. Cluster-5 used five-fold GroupKFold across 54 source-specific clusters: 15 KSS libraries, seven Coastal sites and 32 EukBank studies. Unlike LOPO, it could retain another cluster from the test study in training. The designs therefore represent different prediction settings [14]. Training-fold filtering required at least 5% prevalence and nonzero variance. Forests used 100 trees, balanced class weights and square-root feature sampling with fixed settings and no tuning. A most-frequent-class dummy classifier was refitted within each fold.

Pooled balanced accuracy, the mean recall across true classes, and confusion matrices were the primary summaries; macro-F1 was secondary. Equal-study mean represented-class recall first averages recall over the classes represented in a study and then weights studies equally; a single-class study contributes recall for that class, not a complete three-class discrimination score. Conditional 95% intervals used 1,999 source-study-block resamples of fixed out-of-fold predictions without model refitting. They exclude retraining uncertainty, and no paired test compared the two designs. Complete prediction and performance records are in Table S16.

Supporting RF tasks classified five operational sample types or the three source datasets under 54-unit blocking; these were not replacements for whole-study validation. Exploratory KSS classifiers used leave-one-library-out evaluation for exposure and geographic region. Their settings, prediction records and the Class-level diagnostics are provided in SI Materials and Methods S5 and Tables S16–S18. All primary and supporting results refer to their recorded target, rank, blocking unit and evaluated libraries.

2.9 Statistical Analysis

KSS analyses treated each independently collected solid-substrate sample as one observation; triplicate PCR products did not increase sample size. Comparative analyses used the library and source-group structure specified in Sections 2.1 and 2.6–2.8. Preprocessing, test statistics, permutation restrictions and multiple-testing families are described in Sections 2.3–2.8 and SI Materials and Methods S2–S7. DAA used a BH-adjusted threshold of 0.05 with the additional screens specified above. RF intervals are conditional on fixed out-of-fold predictions, not model refitting. Sample sizes and uncertainty definitions accompany the relevant figures and supplementary tables; R and Python versions are reported with the corresponding main-text and SI methods.

3. Results

3.1 Local assemblages differed in diversity and major eukaryotic contributions

The 15 KSS-18S samples differed markedly in both recovered sequence diversity and dominant taxonomic composition. At a common depth of 20,000 reads, observed richness ranged from 73 ASVs in submerged sediment DGW to 1,034 in air-exposed sediment SWP (Fig. 1A). Shannon diversity ranged

from 2.034 in HFP to 5.486 in SWP, and Simpson diversity from 0.731 to 0.986 in the same samples (Fig. 1B; Figure S5). Pielou's evenness identified a different maximum, ranging from 0.4166 in HFP to 0.7958 in NSP. The two sands illustrate why richness alone did not capture the differences among assemblages: GC and HDP contained 213 and 204 ASVs, respectively, but had evenness values of 0.515 and 0.646. Good's coverage ranged from 0.99350 to 0.99995 at this depth; complete diversity and coverage results are provided in Table S4 and Figure S2.

The taxonomic profiles identify the lineages contributing to these contrasts (Figures S6–S8). Among exposed sediments, SSR contained 55.77% Annelida reads, whereas DBW combined Arthropoda (35.93%) and Chlorophyceae (14.52%). The soil HFS contained 47.85% Basidiomycota. Submerged sediments were also heterogeneous: Arthropoda contributed 79.40% in HFP, while Annelida contributed 53.77% in GKP and 44.95% in SSP. Despite their similar richness, the sands differed in their leading Classes: GC contained 60.18% Mediophyceae, whereas HDP contained 40.74% Annelida and 24.05% Bacillariophyceae. DGW combined a large unresolved-Class fraction (59.93%) with 16.53% Spirotrichea. Thus, differences among samples concerned their dominant constituents as well as their numbers of detected variants.

Retaining the whole-eukaryote scope made these differences visible. Metazoan reads ranged from 1.88% in DGW to 94.11% in HFP and exceeded half of the library in seven samples; fungal reads reached 60.09% in HFS and 10.95% in GJS (Fig. 1C; Figure S3A; Table S11). The operational protist-enriched fraction ranged from 3.525% in HFS to 85.18% in GC and contained 2,747 ASVs and 113,530 reads, or 37.84% of the pooled total.

The pronounced sample-level variation was not accompanied by clear exposed–submerged separation. Air-exposed samples had higher median richness, Shannon diversity and Simpson diversity than submerged samples (401 versus 213 ASVs, 4.024 versus 3.420, and 0.928 versus 0.894), but exact P values were 0.1810, 0.3277 and 0.3277. All six tested alpha-diversity estimators had BH q = 0.3277, with epsilon-squared no greater than 0.077 (Figure S9; Tables S5–S6). Among the 15 KSS samples, the air-exposed versus permanently submerged grouping accounted for 5.86% of Bray–Curtis community-composition variation (pseudo-F = 0.809, exact P = 0.9335; dispersion P = 0.7879), and the groups were interspersed in the contextual ordination (Fig. 1D; Figure S10; Table S7). These estimates describe the exposure contrast in this collection; they do not establish ecological equivalence between exposed and submerged materials.

The submerged-pH comparison remained uncertain (R^2 = 0.2758; exact P = 0.0659; BH q = 0.1317). A separate comparison of three soils with 12 other solids showed an exploratory Genus association (R^2 = 0.1013; P = 0.0269; q = 0.0342 under the comparative-summary correction), but soil was aligned with exposure and location because every soil was air-exposed. Complete alpha-diversity, Genus/Family and dispersion results are reported in SI Results SR2 and Tables S6–S8.

3.2 Taxonomy-inferred lifestyle profiles varied among local assemblages

Taxonomic variation was accompanied by differences in potential trophic affinities, with different profiles at the whole-eukaryote and protist-enriched levels (Figures S3–S4; Table S11). Across the 300,000 whole-eukaryote reads, heterotroph-associated and phototroph-associated annotations accounted for 56.60% and 15.47%, respectively. Much of the former contribution came from animal and fungal sequences. Within the 113,530 protist-enriched reads, phototroph-associated annotations accounted for 39.33%, heterotroph-associated annotations for 9.18%, and variable or insufficiently resolved assignments for 51.48%.

The most conspicuous within-subset contrast concerned phototroph-associated reads, which ranged from 0.56% in DGW to 93.10% in GC (Figure S4B). The GC sand profile was dominated by Mediophyceae, contributing 70.65% of protist-enriched reads, while Bacillariophyceae contributed 49.97% in the other sand, HDP. Chlorophyceae led the protist-enriched profile of the exposed sediment DBW at 61.57% (Figure S3B). Phototroph-associated reads were prominent in GC, HDP and DBW, which included submerged sands and air-exposed sediment. Heterotroph-associated annotations were most prominent within the protist-enriched fraction of HFS, at 49.93%, where Filosa-Sarcomonadea contributed 47.80%. That interpretation concerns 705 retained reads, or 3.525% of the original HFS library; its whole-eukaryote profile was instead fungal-rich. GJS supplied a different protist-enriched soil profile, led by Rigifilida_XX at 28.29%. The whole-library and protist-enriched denominators reveal different components of each sample.

Trophic resolution itself varied among samples. In DGW, the variable or insufficiently resolved category accounted for 93.52% of protist-enriched reads despite identifiable contributions from Spirotrichea (39.93%) and Opalozoa (31.86%). This distinction separates recognition of a lineage from knowledge of its nutritional strategy. The classification rules retained substantial ciliate, chrysophyte and dinoflagellate contributions in variable categories rather than assigning them uniformly to consumption or mixotrophy (Table S11). Only 14 reads from two observed ASVs met the narrow mixotroph-associated rule, approximately 0.01% of the protist-enriched pool. That small annotated fraction describes the reach of the specific rule, not the prevalence of all organisms capable of combining nutritional modes.

Selected parasite- or endobiont-associated lineages added a separate view of possible host relationships (Figure S4C). Apicomplexa reached 2.105% of whole-library reads in HDP and 1.58% in HMC. Phytomyxea reached 0.575% in HMC, Perkinsida 0.42% in HGP, and Pirsoniales 0.10% in GKP (Table S11). These sample-specific values describe separately tracked lineages, not a combined parasite fraction.

3.3 Dominance and recurrence described different aspects of community sharing

The common-depth KSS table contained 4,586 ASVs, none detected in all 15 samples. Only ASV_000018 met the high-recurrence threshold of at least eight samples, occurring in four exposed and four submerged samples (Fig. 2A; Table S9). It contributed 2,237 reads, approximately 0.75% of the pooled total, and was assigned to the PR2 Family placeholder Sphaeropleales_X with unresolved Genus

assignment. Across its eight detections, relative abundance ranged from 0.085% in NSP to 4.465% in DBW. ASV_000018 also exceeded the independent abundance threshold of 0.2% and was therefore one of the 66 dominant ASVs.

The 66 abundance-selected ASVs collectively accounted for 171,060 reads, or 57.02% of the pooled total (Fig. 2B; Table S10). High pooled abundance did not imply widespread occurrence. ASV_000001, an annelid assignment, contributed 4.86% and occurred in five samples; ASV_000004 (*Cypria*) contributed 4.13% in three; ASV_000003 (*Strandesia*) contributed 3.94% in four; and the annelid ASV_000002 contributed 3.92% in four. ASV_000005, assigned to Thalassiosiraceae, contributed 3.28% despite being detected only in GC. Other dominant sequences were more recurrent: the *Desmodesmus* ASV_000027 occurred in seven samples and the ostracod ASV_000010 in six. The abundance-selected set thus combined recurrent sequences with large contributions concentrated in particular samples rather than defining a ubiquitous core.

The individual ASV profiles also connect the broad taxonomic contrasts to particular recovered sequences (Fig. 2C; Figure S11). DGW included the dominant Oxytrichidae ASV_000014 and unresolved ASV_000008 and ASV_000012. HDP contained ASV_000038, assigned within Litostomatea to Kentrophyllidae. Agaricomycete ASV_000013 and ASV_000015 contributed prominently in HFS, and the *Micronuclearia* ASV_000020 in GJS. Community sharing changed with taxonomic resolution. Chrysophyceae and Chytridiomycota occurred at Class level in all 15 samples, even though no constituent ASV was ubiquitous (Table S12). The V4 sequence display likewise places related variants alongside different sample profiles (Figure S11).

Distribution breadth distinguished frequent detection from an even distribution of reads. Among 94 eligible Class-level features, standardized sample-based Levins breadth ranged from 0.0014 for Echinamoebida to 0.5206 for Oligohymenophorea (Figure S12; Table S12). Annelida and Arthropoda contributed 16.97% and 15.33% of pooled reads and occurred in ten and 13 samples, but had breadth values of 0.3630 and 0.2909. Their frequent occurrence thus coexisted with unequal contributions among samples. Category-based breadth and abundance-weighted pH profiles apply a different weighting to the available observations (Figure S13; Table S2). The Coastal pH profile is specifically a site-context summary: site means include all retained Coastal libraries, whereas pH was measured in spring sea-sand.

3.4 Local heterogeneity coexisted with broader sample-type-associated structure

The 140-library comparison, comprising 15 KSS-18S, 75 Coastal and 50 selected EukBank libraries, revealed broader sample-type-associated structure at both taxonomic ranks. After dataset adjustment, sample type accounted for 9.77% of Genus and 12.04% of Family Bray–Curtis variation (pseudo-F = 3.83 and 4.96; P = 0.0001 for both; Fig. 3; Table S8). The full-model residual fractions were 84.71% and 80.74%, respectively. The dataset and sample-type R^2 values are marginal term estimates and need not sum with the residual fraction to 100%; sample type explained a modest component of variation rather

than separating all libraries into discrete groups. The P values refer to permutations restricted within the three datasets; these restrictions did not fully represent finer Coastal-site and EukBank-study dependencies. The dataset-term P value was not interpreted because dataset membership was not exchangeable under that restriction.

The finer-block sensitivity retained a within-block sample-type association after conditioning on 46 blocks. Fifteen blocks contained more than one sample type, yielding two estimable degrees of freedom and 92 residual degrees of freedom. The tested component explained 5.81% of total Genus and 6.40% of total Family variation (pseudo-F = 6.78 and 8.03; P = 0.0001 and two-rank Holm-adjusted P = 0.0002 for both). These values are not replacements for the dataset-adjusted estimates because the conditioning and estimable contrast space differ. The same restricted permutations also detected dispersion differences (P = 0.0001 at both ranks; Table S8). Thus, differentiation remained evident under finer blocking, but a centroid-only interpretation is not supported. The within-block information concerns land soil versus other KSS solids and marine water versus marine sediment; it does not independently test freshwater or land-sediment–marine-water differences within source studies.

Bray–Curtis PCoA and CLR-PCA showed variation among and within the operational sample types, while the leading assigned Genus and Family profiles identified their taxonomic contributions (Fig. 3; Figure S14). Bray–Curtis PCoA represents dissimilarities in relative composition, whereas CLR-PCA represents variation in log ratios; their axis percentages are not interchangeable. Assignment coverage also matters because these comparisons use assigned reads. In the KSS common-depth summary, 33.28% of reads were unassigned at Genus and 11.63% at Family (Figure S15; Table S3).

The Coastal reference provides a particularly useful within-study comparison. Its 25 site–season–sample-type combinations each had three independent biological replicates, yielding 75 libraries rather than 25 averaged observations. In the present analysis of the DADA2-reprocessed tables, after site and season adjustment, sample type accounted for 10.29% of Genus and 12.71% of Family variation (P = 0.0001), while site terms accounted for 26.57% and 24.25% (Table S8). Thus, a sand–water contrast coexisted with substantial variation among sites under a common sampling design.

EukBank broadened the environmental comparison but supplied fewer within-study habitat contrasts. Sample-type estimates were 18.06% at Genus and 20.94% at Family, yet only two of the 32 selected source studies crossed habitat categories. Study-restricted testing admitted 18 distinct allocations and gave exact P = 0.0556 at both ranks, compared with Monte Carlo P = 0.0522 (Table S8). Uneven sequencing effort can influence community distances [28]. We therefore examined the sensitivity of the distance-based result to read depth. Genus sample-type R^2 values were 9.77% for raw relative composition, 9.73% for mean 14k distances and 9.71% for mean 10k distances; the corresponding Family values were 12.04%, 12.01% and 11.99% (Figure S16; Table S13). All six tests had P = q = 0.0001, and distance correlations were at least 0.99989. In separate raw-distance models, $\log_{10}$ read count explained 0.71% and 0.77% of variation, with P = 0.3544 and 0.2225. The reported association therefore

changed little under these depth treatments. These checks concern the specified distances and permutation designs, not the independent validity of CLR-PCA, DAA or classification.

3.5 Family candidates represented exploratory between-study associations

The pooled DAA identified a small, qualified set of Family associations rather than a general taxonomic signature for every sample type. ANCOM-BC2 models without source-cluster random intercepts retained 19 pseudo-count-robust Family and 12 Genus terms. With those intercepts, three Family terms and no Genus terms remained (Fig. 4; Figures S17–S18; Tables S14–S15). These were taxon-by-contrast coefficients, not necessarily distinct taxa across the full results. All three retained Family terms concerned marine water relative to land sediment and had negative coefficients. No finite robust Family coefficient remained for land soil, freshwater or marine sediment relative to land sediment. Structural-zero flags describe a different property of the profiles and are reported separately from the finite coefficients in Fig. 4, Table S14 and SI Results SR6.

Oxytrichidae had a $\log_2$ coefficient of − 7.65 (BH q = 0.0258), Sandonidae − 8.52 (q = 0.0051), and Litostomatea_XX − 6.47 (q = 0.0496). All three coefficients were estimable and passed pseudo-count sensitivity screening, but the Litostomatea_XX random-effect fit was singular; Oxytrichidae and Sandonidae had nonsingular fits (Table S15). The evidence comprises two nonsingular exploratory associations and one association with a singular random-effect fit; these fitting diagnostics do not by themselves establish ecological reproducibility. The marine-water–land-sediment contrast was not represented within any source study in the assembled comparison; these are therefore between-study associations, not within-study habitat contrasts (Table S3). The random intercept does not supply the missing within-study comparison or establish a habitat effect independent of source-study differences. The negative coefficients indicate lower model-estimated abundance in marine water than in land sediment under the specified model, not calibrated differences in cell numbers or biomass [29].

The Litostomatea ASV_000038 in HDP belongs to Kentrophyllidae, not to the PR2 Family placeholder Litostomatea_XX, and cannot corroborate that Family-level term.

Within-dataset results reinforce the need to specify the contrast. No KSS ASV, Genus or Family ANCOM-BC2 term passed the robust screen; the Class analysis identified one nominal exposure term and two nominal submerged-pH terms, but none passed that additional screen (Figure S19; Table S5). Coastal retained 73 Family and 67 Genus terms without site random intercepts, compared with 46 and 42 with them. EukBank retained three Family terms without study random intercepts and none with them (Table S14). These results concern different contrasts and do not constitute independent repetitions of the pooled land-sediment–marine-water comparison.

ALDEx2 supplied within-dataset sensitivity comparisons. It retained no KSS ASV among 182 tested for exposure and no Family among 211 tested for soil versus other solids. One Genus, *Kraken*, was retained among 277 in the latter contrast, with expected Welch BH q = 0.0444, a between-group $\log_2$ CLR

difference of 8.133 and a standardized effect of 2.354 (Figures S20–S21; Table S14). *Kraken* is a scale-bearing filose cercozoan [32], but this isolated result remained a method-specific association without a corresponding robust KSS ANCOM-BC2 result. Coastal ALDEx2 comparisons retained 116 of 249 Families and 118 of 275 Genera; complete EukBank contrasts are given in SI Results SR6. In the overlapping EukBank subset, ALDEx2 estimates for marine water minus land sediment were negative for all three Family candidates, but none had BH $q < 0.05$ (Table S14). These comparisons had no source-study random effects and were not independent validation; no ALDEx2 analysis was performed on the pooled 140-library design.

3.6 Community profiles transferred unevenly to withheld source studies

Whole-profile classification showed that broader differentiation in the assembled data did not translate into uniformly successful cross-study prediction. For the matched target of sediment/sand, soil and water, LOPO balanced accuracy was 0.4786 at Genus and 0.4864 at Family, with conditional 95% intervals of 0.3648–0.7038 and 0.3686–0.7282 (Fig. 5; Table S16). Cluster-5 gave 0.7384 (0.5346–0.8422) and 0.7187 (0.5139–0.8432), compared with dummy balanced accuracy of 0.3333. The designs address different prediction settings: LOPO withholds one of 34 entire source studies, whereas Cluster-5 divides 54 source-specific clusters among five folds and can retain other clusters from a test study in training. The comparison is descriptive; no paired significance test was performed, and the intervals resample fixed predictions without model refitting.

Both ranks correctly assigned 56 of 71 sediment/sand libraries. Only four of 13 soils were correctly classified at Genus and five at Family, while water recall was 19 of 56 and 16 of 56. Most errors for soil and water accumulated in sediment/sand: nine and eight soil libraries, and 37 and 40 water libraries, respectively, received that label (Fig. 5B–C; Table S16). Thus, relatively high sediment/sand recall coexisted with frequent assignment of other materials to the same class; it did not define a unique solid-substrate signature. Overall LOPO accuracy was 0.5643 and 0.5500, compared with 0.8143 at both ranks under Cluster-5 and dummy accuracy of 0.5071.

Equal-study summaries changed the weighting of performance rather than the prediction task. LOPO equal-study mean represented-class recall was 0.5849 at Genus and 0.6139 at Family, compared with 0.6346 and 0.6301 for Cluster-5 and 0.3529 for the dummy classifier (Table S16). Thirty of the 34 studies contained only one true class, so their contribution was recall for that represented class, not within-study three-class discrimination. For the two larger withheld sources, KSS LOPO balanced accuracy was 0.4167 and 0.4583, and Coastal values were 0.4690 and 0.4135. These summaries show variation among evaluation units while retaining the consequences of unequal class and study representation.

Supporting tasks localized other sources of discrimination without replacing whole-study validation. Under 54-unit blocking, the five-sample-type task reached balanced accuracy of 0.6505 at Genus and 0.6527 at Family (Figures S22–S23; Table S17). Within KSS, exploratory leave-one-library-out exposure classification gave balanced accuracy of 0.5833 and 0.5556. Geographic classification assigned every

sample to Jeju, giving accuracy of 0.80 but balanced accuracy of 0.50 (Figure S24; Table S18). Class-level RF diagnostics are provided separately in Figure S25.

4. Discussion

4.1 Local heterogeneity and environmental interpretation

Because the retained protist-enriched fraction varied strongly among KSS libraries, a protist-only inventory would describe very different fractions of the original samples. These proportions identify contrasting contributions to the recovered sequence pool, not differences of the same magnitude in organism abundance or biomass, because rDNA copy number and assay recovery vary among eukaryotes [10, 12, 13].

These observations suggest that broad exposure labels do not summarize all biologically relevant variation among the sampled materials. Environmental associations reported for soil and lake protists involve moisture, edaphic properties, trophic state and watershed context, with responses that differ among lineages [2, 3, 4]. In the soil survey of Oliverio et al. [2], mean annual precipitation was the leading environmental correlate of community composition. That result makes water availability a relevant biological question, but air exposure versus permanent submergence does not measure the same gradient. Evidence from Baltic plankton highlights the importance of measured abiotic gradients in interpreting protist diversity, community composition and realized niche breadth [33]. The weak exposed–submerged separation in KSS should therefore not be interpreted as evidence of weak environmental structuring, because these categories do not resolve the underlying abiotic gradients.

4.2 Potential trophic affinities and their biological limits

Oliverio et al. [2] reported mean consumer, phototroph and parasite fractions of 85%, 10% and 5% across 180 protist-only soil profiles. Direct numerical comparison with KSS is limited by differences in organism inclusion, weighting and trophic definitions. Our whole-eukaryote profiles included animals and fungi, while the pooled protist-enriched profile was read-weighted. Heterotroph-associated annotations did not define a consumer guild, and parasite-associated lineages were tracked separately. The KSS profiles raise food-web questions about possible photosynthetic inputs in algal-rich GC, consumption by phagotroph-associated lineages and host dependence in separately tracked groups. These are potential ecological roles, not measured processes or uniform properties of a substrate.

Oliverio et al. [2] likewise examined parasite-associated groups separately to resolve their distributions, although the groups and environmental comparison differed. Host identity, activity and infection status remain questions for organism-linked observations. Animal reads in the same bulk samples provide possible host context but do not demonstrate a host–parasite link.

4.3 Dominance, recurrence and taxonomic resolution

Shared higher-rank assignments can therefore coexist with differences in the particular sequences recovered. Oliverio et al. [2] also aggregated recurrent ASVs into phylogenetic clades because individual variants were sparsely shared among soils. That approach addresses a different sequence set from our abundance-selected ASVs, but reinforces the importance of stating the level at which sharing is evaluated. Soil–freshwater comparisons using V9 sequence units have also found limited sequence overlap despite the occurrence of related taxa in both environments [5]. That study offers a conceptual comparison, not a numerical expectation for KSS V4 recurrence. Short sequences can merge distinctions retained by longer 18S sequences, and ASVs are not equivalent to organismal species [7, 11]. The present patterns therefore identify sequence-level heterogeneity beneath shared higher taxa without demonstrating species replacement or reconstructing whole-community evolutionary history. Category-based breadth and abundance-weighted pH profiles summarize distributions across the sampled categories or pH records, rather than physiological tolerance or generalist–specialist status. This distinction is useful for choosing which lineages and environmental contexts merit more closely matched sampling.

All recurrence statements are conditional on the 20,000-read table and its detection criterion. Rare variants may be missed, and the historical comparison of denoising strategies changed the recurrence distribution (Additional Data Figure AD1). That comparison does not replace the detection criterion of the common-depth table or constitute a repeated-subsampling test of recurrence. Failure to detect any ASV in all samples therefore does not establish that no organism is widespread. Nor does sparse sharing test movement restriction or temporal stability, because locations were sampled once.

4.4 What the external comparisons add

The two ranks therefore represent different fractions of the recovered inventory as well as different biological resolution. Shared rank definitions and coverage records document the analyzed representation but do not establish identical taxon recovery across studies. Axis percentages and the distinct unrarefied comparative coverage summaries are retained in SI Results SR5.

In the Coastal reference, water was size-fractionated and sea-sand was not, so the sand–water contrast combines environmental context with different organism-recovery procedures; their contributions were not separated. This is relevant to the interpretation of KSS: local heterogeneity need not preclude a broader sample-type association. It does not, however, make the coastal estimates applicable to unreplicated KSS locations or identify a single environmental mechanism shared by all materials.

The coarse resolution of the EukBank study-restricted test limits its inferential contribution; it is not evidence that the communities lack habitat-associated differences. Unrestricted dispersion diagnostics and the later matched restricted checks used different geometries and restrictions and are reported separately in SI Results SR5. The references consequently contribute different kinds of evidence: Coastal supplies a replicated within-study contrast, whereas the selected EukBank libraries extend the contexts represented with limited independent within-study information for habitat testing.

The broader differentiation is compatible with comparisons of protist assemblages across soil, freshwater and marine environments [1], although the present study retained the wider eukaryotic inventory. It also addresses a different question from the KSS exposure comparison: operational sample types span environmental contexts not represented by air exposure alone. The external comparisons therefore place the heterogeneous KSS profiles in a wider sampling context rather than independently confirming a common environmental effect.

4.5 Interpreting the exploratory Family associations

These candidates connect the comparative result to identifiable components of the local inventory. Oxytrichidae includes the dominant ASV_000014 in DGW, providing a concrete example of a candidate Family contributing prominently to a KSS sediment (Table S10). Sandonidae are prominent in soil protist inventories, placing their association with the land-sediment side of the contrast in a relevant terrestrial context [2]. That study also assigned different Sandonidae clades to dry and cold environmental clusters, showing that a shared Family label can contain distinct distributional responses. The present Family association therefore identifies a lineage group for closer study, rather than a uniform habitat requirement of all its members. Neither the local ASV contribution nor the soil comparison independently validates the pooled coefficient.

The biological question suggested by these findings is whether better-resolved members of the candidate lineages show comparable associations in independently sampled land sediment and marine water.

The smaller retained set after source-cluster adjustment shows that candidate selection depended on model specification. It does not establish why each term crossed or failed a threshold: random effects can change uncertainty and estimability as well as point estimates. Rank-dropped fits were ineligible, and the singular Litostomatea_XX fit remains distinguishable from the two nonsingular candidates (Tables S14–S15).

The methods and ranks should therefore be read as complementary, not interchangeable. ALDEx2 expresses between-group differences relative to a geometric-mean denominator and reports a separate standardized effect, whereas ANCOM-BC2 estimates bias-corrected log-abundance contrasts under its sampling and taxon-bias assumptions [29, 30]. These distinctions are important for relative amplicon data and their dependence structure [34, 35]. Likewise, the retention of Family but not Genus terms does not identify an intrinsically superior ecological rank: aggregation and assignment coverage also differ. The analysis identified exploratory lineage–sample-type associations for independent testing.

4.6 Prediction depends on the withheld source and class representation

The representation available during training helps frame this asymmetry. Withholding Coastal removes 75 libraries, including 36 of the 56 water libraries, leaving water examples from EukBank. Withholding KSS removes three of the 13 soil libraries. Moreover, the three-class water label combines freshwater

and marine materials that behaved differently in the supporting five-class task: marine-water recall was 1.00 but freshwater recall was 0.10 (Figure S22; Table S17). These properties offer plausible explanations for uneven transfer, without distinguishing their effects from source-associated biological or laboratory differences. Dataset classification accuracy of 0.8357 at both ranks shows that source-associated information remained in the profiles, but cannot partition that information into ecological and technical components. The within-KSS geographic result illustrates why class balance and the evaluated contrast must accompany accuracy: a correct assignment rate can largely reflect the commonest label.

The weaker and asymmetric whole-study transfer is compatible with differences among habitats that also contain substantial site- and study-associated variation. It does not negate those associations, but it limits their immediate use for recognizing an unknown sample's context. Matching validation to intended prediction is essential for structured ecological data [14]. Independent observations are needed to determine which associations recur beyond this collection.

4.7 Limitations

KSS contains independently collected samples, but each location was visited once and location, substrate, exposure and sampling year were not fully crossed. The external references extend the contexts represented without providing repeat observations at those locations. The pooled community comparisons, the ANCOM-BC2 models used to identify candidate Family associations, and the RF classifiers also remain conditional on their source structure and sampling coverage. These features constrain environmental attribution and generalization, rather than the description of the recovered assemblages.

These constraints also apply to the environmental summaries. KSS numerical pH records combine point values and ranges, Coastal pH describes spring sea-sand site context, and the selected EukBank libraries have no pH records. A matched multivariable environmental design is therefore unavailable across the comparison. The EukBank subset is selective rather than a comprehensive census; its exact source snapshot and retrieval date remain unresolved. Available selection records document the subset without reconstructing missing provenance.

4.8 Directions for further ecological research

Repeated sampling could test whether the contrasting local profiles recur under comparable conditions. Repeated collections of the same materials across geographic regions and seasons, with matched pH, moisture, light, temperature, redox, organic matter and grain-size measurements, could distinguish repeatable associations from differences specific to a location or sampling occasion. Geographic coverage and environmental coverage should be considered together. Oliverio et al. [2] identified uneven representation of regions and biomes in their global soil survey and argued that sampling poorly represented settings would improve the resolution of protist distributions. That spatial objective is distinct from, but complementary to, improving taxonomic identification. In KSS, the pH categories support categorical comparisons and the numerical points or ranges support descriptive profiles; they do not substitute for matched multivariable measurements. For the Family associations, independent

studies containing both land sediment and marine water under a common protocol would be particularly informative and would provide a clearer basis for evaluating cross-study prediction.

A second question concerns the organisms represented by prominent or unresolved sequences. The large unresolved fraction in DGW, the animal-rich libraries and the differing protist-enriched profiles identify concrete priorities for taxonomic follow-up. Longer 18S sequences and linked references could improve placement of V4 variants; complementary cytochrome c oxidase subunit I (COI) gene and ribosomal internal transcribed spacer (ITS) assays could refine animal and fungal components, respectively [36, 37]. Microscopy or imaging of matched samples could then relate sequence profiles to identifiable organisms. The PCR-negative control showed no visible gel band but was not sequenced, so sequence-level contamination was not assessed in this survey. Sequenced blanks in subsequent collections would help evaluate low-frequency detections alongside improved taxonomic resolution.

The lifestyle profiles make the follow-up questions sample-specific. Algal cell identification and measurements of light-dependent activity would help resolve the ecological contribution suggested by the high phototroph-associated read fraction in sample GC. Finer identification of the conspicuous but trophically unresolved DGW lineages would address a different gap, while host-linked observations are needed for the Apicomplexa signals in HDP and HMC. The complementary shotgun comparison in Oliverio et al. [2] illustrates one way to check the identities of abundant protist lineages; it does not validate the KSS annotations. Matched shotgun data could also extend these inventories to protein-coding genes and metabolic potential, although eukaryotic gene discovery and incomplete references remain important considerations [38]. Contemporaneous RNA, protein or physiological evidence would then be needed to connect organisms and genes to processes.

Paired prokaryotic (bacterial and archaeal) and eukaryotic amplicon profiles, shotgun metagenomes and matched abiotic measurements across regions and seasons could nominate potential prey–predator relationships in sediment, sand and soil. Co-occurrence and gene content would generate hypotheses; repeated sampling and direct feeding observations would be needed to test them. This extends the protist–prokaryote questions raised by Oliverio et al. [2], while the proposed interactions and their dynamics remain untested by the present survey.

5. Conclusions

The KSS survey revealed marked variation in eukaryotic diversity and dominant animal, fungal and algal contributions among sediment, sand and soil samples. Protist-enriched profiles added sample-specific contrasts in potential trophic affinity, not measurements of ecological activity. A small abundance-selected set accounted for more than half of the recovered reads, yet no ASV was ubiquitous; shared higher-rank taxa coexisted with sequence-level heterogeneity. External references revealed modest sample-type-associated structure, accompanied by dispersion differences under finer blocking. Oxytrichidae and Sandonidae showed exploratory between-study Family associations in models with nonsingular random-effect fits. The Litostomatea_XX model had a singular random-effect fit. Whole-

study prediction transferred unevenly, especially for soil and water. The primary contribution is a sample-resolved inventory that distinguishes local dominance from recurrence across solid substrates. External comparisons extend its context and nominate taxa and contrasts for independent testing, but do not establish habitat effects independent of source-study differences or validated ecological markers.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported by grants from the National Research Foundation of Korea (no. 2020R1I1A3062110 to SJP) and the Korea Environmental Industry & Technology Institute (KEITI) through the Soil Carbon Sequestration Project, funded by the Ministry of Climate, Energy and Environment (MCEE) (RS-2026-25504681 to SJK).

Author Contribution

MK and SJP conceived and designed the study. MK, SJK and SJP performed the experiments and analyzed the data. MK and SJP contributed to conceptualization, data curation and validation. SJP prepared the original draft, and MK, SJK and SJP reviewed and edited the manuscript.

Acknowledgement

We are grateful to Professor Meng Li (Shenzhen University) for insightful discussions and constructive feedback on the overall framing, interpretation, and discussion of this manuscript.

Data Availability

The raw 18S rRNA gene amplicon reads generated in this study have been deposited in the NCBI Sequence Read Archive under BioProject PRJNA1459660 and SRA accessions SRX33162133–SRX33162148 [39]. Selected processed data for KSS-18S, Coastal and EukBank, together with analysis scripts, are publicly available in the KSS-18S directory of the microbial-ecology-analysis GitHub repository [40].

References

1. Singer D, Seppey CVW, Lentendu G, Dunthorn M, Bass D, Belbahri L, et al. Protist taxonomic and functional diversity in soil, freshwater and marine ecosystems. Environ Int. 2021;146:106262. https://doi.org/10.1016/j.envint.2020.106262.
2. Oliverio AM, Geisen S, Delgado-Baquerizo M, Maestre FT, Turner BL, Fierer N. The global-scale distributions of soil protists and their contributions to belowground systems. Sci Adv. 2020;6:eaax8787. https://doi.org/10.1126/sciadv.aax8787.
3. Bates ST, Clemente JC, Flores GE, Walters WA, Parfrey LW, Knight R, et al. Global biogeography of highly diverse protistan communities in soil. ISME J. 2013;7:652–9. https://doi.org/10.1038/ismej.2012.147.
4. Garner RE, Kraemer SA, Onana VE, Huot Y, Gregory-Eaves I, Walsh DA. Protist diversity and metabolic strategy in freshwater lakes are shaped by trophic state and watershed land use on a continental scale. mSystems. 2022;7:e0031622. https://doi.org/10.1128/msystems.00316-22.
5. Sieber G, Beisser D, Bock C, Boenigk J. Protistan and fungal diversity in soils and freshwater lakes are substantially different. Sci Rep. 2020;10:20025. https://doi.org/10.1038/s41598-020-77045-7.
6. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13:581–3. https://doi.org/10.1038/nmeth.3869.
7. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 2017;11:2639–43. https://doi.org/10.1038/ismej.2017.119.
8. Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, et al. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. Nucleic Acids Res. 2013;41:D597–604. https://doi.org/10.1093/nar/gks1160.
9. Vaulot D, Sim CWH, Ong D, Teo B, Biwer C, Jamy M, et al. metaPR2: A database of eukaryotic 18S rRNA metabarcodes with an emphasis on protists. Mol Ecol Resour. 2022;22:3188–201. https://doi.org/10.1111/1755-0998.13674.
10. Prokopowich CD, Gregory TR, Crease TJ. The correlation between rDNA copy number and genome size in eukaryotes. Genome. 2003;46:48–50. https://doi.org/10.1139/g02-103.
11. Lie AAY, Liu Z, Hu SK, Jones AC, Kim DY, Countway PD, et al. Investigating microbial eukaryotic diversity from a global census: insights from a comparison of pyrotag and full-length sequences of 18S rRNA genes. Appl Environ Microbiol. 2014;80:4363–73. https://doi.org/10.1128/aem.00057-14.

12. Choi J, Park JS. Comparative analyses of the V4 and V9 regions of 18S rDNA for the extant eukaryotic community using the Illumina platform. Sci Rep. 2020;10:6519. https://doi.org/10.1038/s41598-020-63561-z.
13. Han Y, Guo C, Guan X, McMinn A, Liu L, Zheng G, et al. Comparison of deep-sea picoeukaryotic composition estimated from the V4 and V9 regions of 18S rRNA gene with a focus on the hadal zone of the Mariana Trench. Microb Ecol. 2022;83:34–47. https://doi.org/10.1007/s00248-021-01747-2.
14. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017;40:913–29. https://doi.org/10.1111/ecog.02881.
15. Kim M, Cha IT, Park SJ. Water-sand contrasts and site-associated variation in bacterial and eukaryotic communities along the Korean coast. Microbiol Spectr. 2026;e02092–26. https://doi.org/10.1128/spectrum.02092-26.
16. Berney C, Mahé F, Henry N, Lara E, de Vargas C, EukBank consortium. EukBank 18S V4 dataset [dataset]. Zenodo; 2023. https://doi.org/10.5281/zenodo.7804946.
17. International Organization for Standardization. Soil, Treated Biowaste and Sludge—Determination of pH. ISO 10390:2021. Geneva: International Organization for Standardization; 2021. https://www.iso.org/standard/75243.html.
18. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B. 1995;57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
19. Bray JR, Curtis JT. An ordination of the upland forest communities of southern Wisconsin. Ecol Monogr. 1957;27:325–49. https://doi.org/10.2307/1942268.
20. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 2001;26:32–46. https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x.
21. Anderson MJ. Distance-based tests for homogeneity of multivariate dispersions. Biometrics. 2006;62:245–53. https://doi.org/10.1111/j.1541-0420.2005.00440.x.
22. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS ONE. 2010;5:e9490. https://doi.org/10.1371/journal.pone.0009490.
23. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80. https://doi.org/10.1093/molbev/mst010.
24. Adl SM, Bass D, Lane CE, Lukeš J, Schoch CL, Smirnov A, et al. Revisions to the classification, nomenclature, and diversity of eukaryotes. J Eukaryot Microbiol. 2019;66:4–119. https://doi.org/10.1111/jeu.12691.
25. Levins R. Evolution in Changing Environments: Some Theoretical Explorations. Princeton: Princeton University Press; 1968. https://doi.org/10.2307/j.ctvx5wbbh.
26. Lingoes JC. Some boundary conditions for a monotone analysis of symmetric matrices. Psychometrika. 1971;36:195–203. https://doi.org/10.1007/bf02291398.

27. Aitchison J, Stat Soc B. 1982;44:139–60. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x.
28. Schloss PD. mSystems. 2026;e00930–26. https://doi.org/10.1128/msystems.00930-26. Rarefaction is better than robust Aitchison PCA and other compositional data analysis methods at controlling for uneven sequencing effort.
29. Lin H, Peddada SD. Multigroup analysis of compositions of microbiomes with covariate adjustments and repeated measures. Nat Methods. 2024;21:83–91. https://doi.org/10.1038/s41592-023-02092-7.
30. Fernandes AD, Reid JNS, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome. 2014;2:15. https://doi.org/10.1186/2049-2618-2-15.
31. Breiman L. Random forests. Mach Learn. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.
32. Dumack K, Mylnikov AP, Bonkowski M. Evolutionary relationship of the scale-bearing Kraken (incertae sedis, Monadofilosa, Cercozoa, Rhizaria): combining ultrastructure data and a two-gene phylogeny. Protist. 2017;168:362–73. https://doi.org/10.1016/j.protis.2017.04.004.
33. Jurdzinski KT, Latz MAC, Torstensson A, Brugel S, Hedblom M, Hu YOO, et al. DNA-based monitoring uncovers contrasting salinity-driven diversity patterns of bacterial and protist plankton in the Baltic Sea area. Commun Biol. 2026;9:1308. https://doi.org/10.1038/s42003-026-10977-4.
34. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Front Microbiol. 2017;8:2224. https://doi.org/10.3389/fmicb.2017.02224.
35. Boshuizen HC, te Beest DE. Pitfalls in the statistical analysis of microbiome amplicon sequencing data. Mol Ecol Resour. 2023;23:539–48. https://doi.org/10.1111/1755-0998.13730.
36. Hebert PDN, Cywinska A, Ball SL, deWaard JR. Biological identifications through DNA barcodes. Proc R Soc Lond B Biol Sci. 2003;270:313–21. https://doi.org/10.1098/rspb.2002.2218.
37. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc Natl Acad Sci USA. 2012;109:6241–6. https://doi.org/10.1073/pnas.1117018109.
38. Levy Karin E, Mirdita M, Söding J. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. Microbiome. 2020;8:48. https://doi.org/10.1186/s40168-020-00808-x.
39. National Center for Biotechnology Information. BioProject PRJNA1459660. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1459660. Accessed 23 Oct 2026.
40. KSS-18S. analysis code and selected data. microbial-ecology-analysis GitHub repository. https://github.com/sjparkmk/microbial-ecology-analysis/tree/main/KSS-18S. Accessed 23 Oct 2026.

Figures

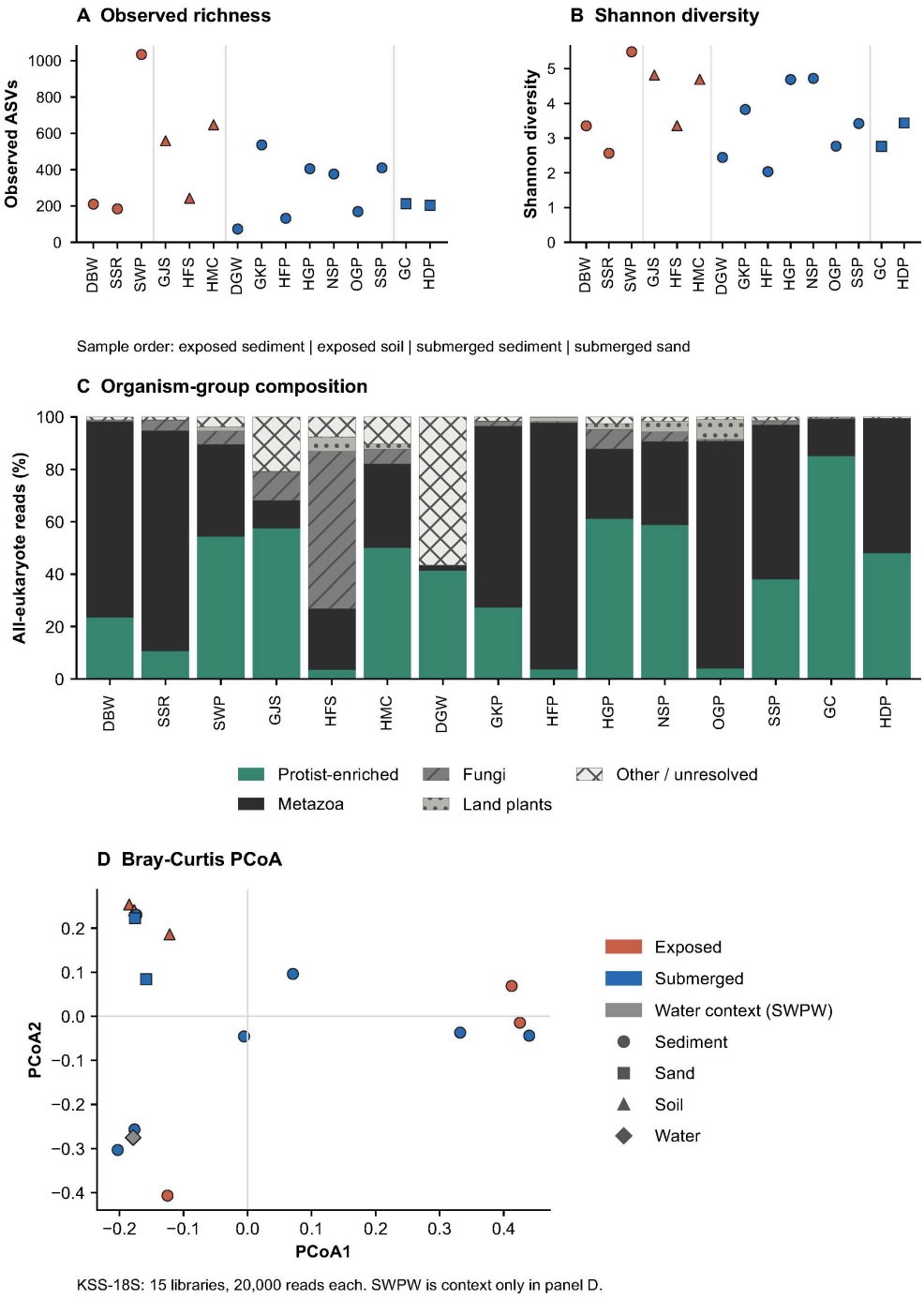


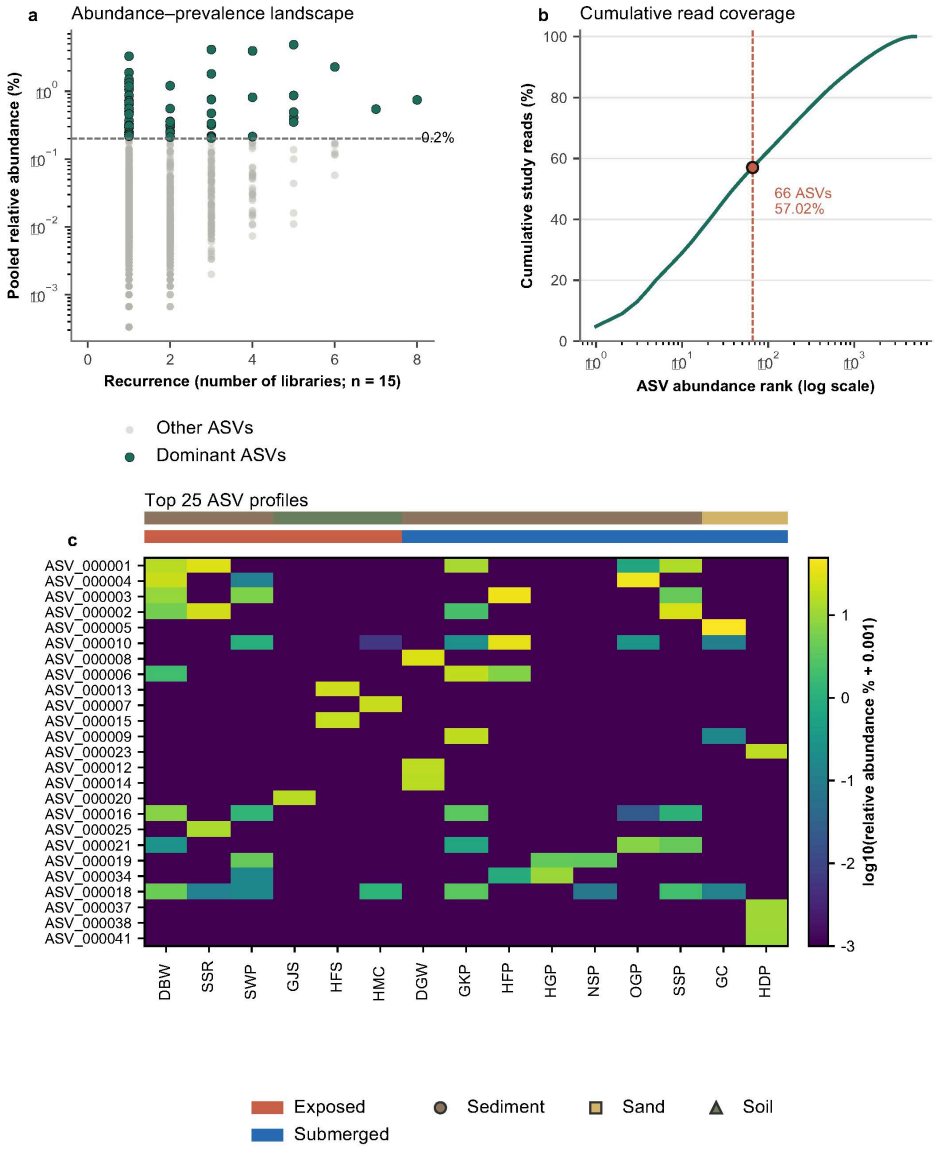
Figure 1

Diversity and composition of KSS-18S

(A) Observed amplicon sequence variants (ASVs) and (B) Shannon diversity at 20,000 reads per KSS-18S sample. Colour distinguishes exposure and shape distinguishes sediment, sand and soil. Samples are ordered as exposed sediment, exposed soil, submerged sediment and submerged sand. Simpson

diversity and Pielou's evenness are included in Figure S5. (C) Organism-group read percentages for the 15 KSS libraries, using the Table S11 annotations and all 20,000 reads per library as denominator. Protist-enriched, metazoan, fungal, land-plant and other/unresolved contributions sum to 100%; these are not measured feeding activities. Division-level profiles are shown in Figures S6–S8. (D) Bray–Curtis principal coordinates analysis (PCoA). SWPW is a contextual water-column library shown only in panel D and excluded from the 15-sample statistical tests. Panel D uses the 16-library ordination; inference for exposure and pH uses the 15 solid-substrate libraries. Exposure and substrate are encoded separately. Read fractions are not cell or biomass fractions. Exact diversity and group-test results are provided in Tables S4–S7.

Figure 2. ASV abundance and recurrence



Aligned tracks: exposure (lower), substrate (upper). Full ASV taxonomy is retained in the source table.

Figure 2

Dominant ASVs and occurrence across samples

(a) Pooled read abundance versus recurrence, expressed as the number of the 15 common-depth libraries with a detection. Detection requires at least one read, and dominant amplicon sequence variants (ASVs) are those contributing at least 0.2% of 300,000 reads (dashed line). (b) Cumulative read

contribution by amplicon sequence variant (ASV) abundance rank; the 66 dominant ASVs account for 57.02%. (c) Sample profiles of the 25 most abundant ASVs on a $\log_{10}$(relative abundance % + 0.001) scale, with exposure (lower) and substrate (upper) tracks. Dominance is defined independently of recurrence; ASV_000018 meets both the dominant and high-recurrence criteria. No ASV was detected in all 15 samples. Complete recurrence and selected-ASV records are in Tables S9–S10.

Figure 3. Comparative community ordination

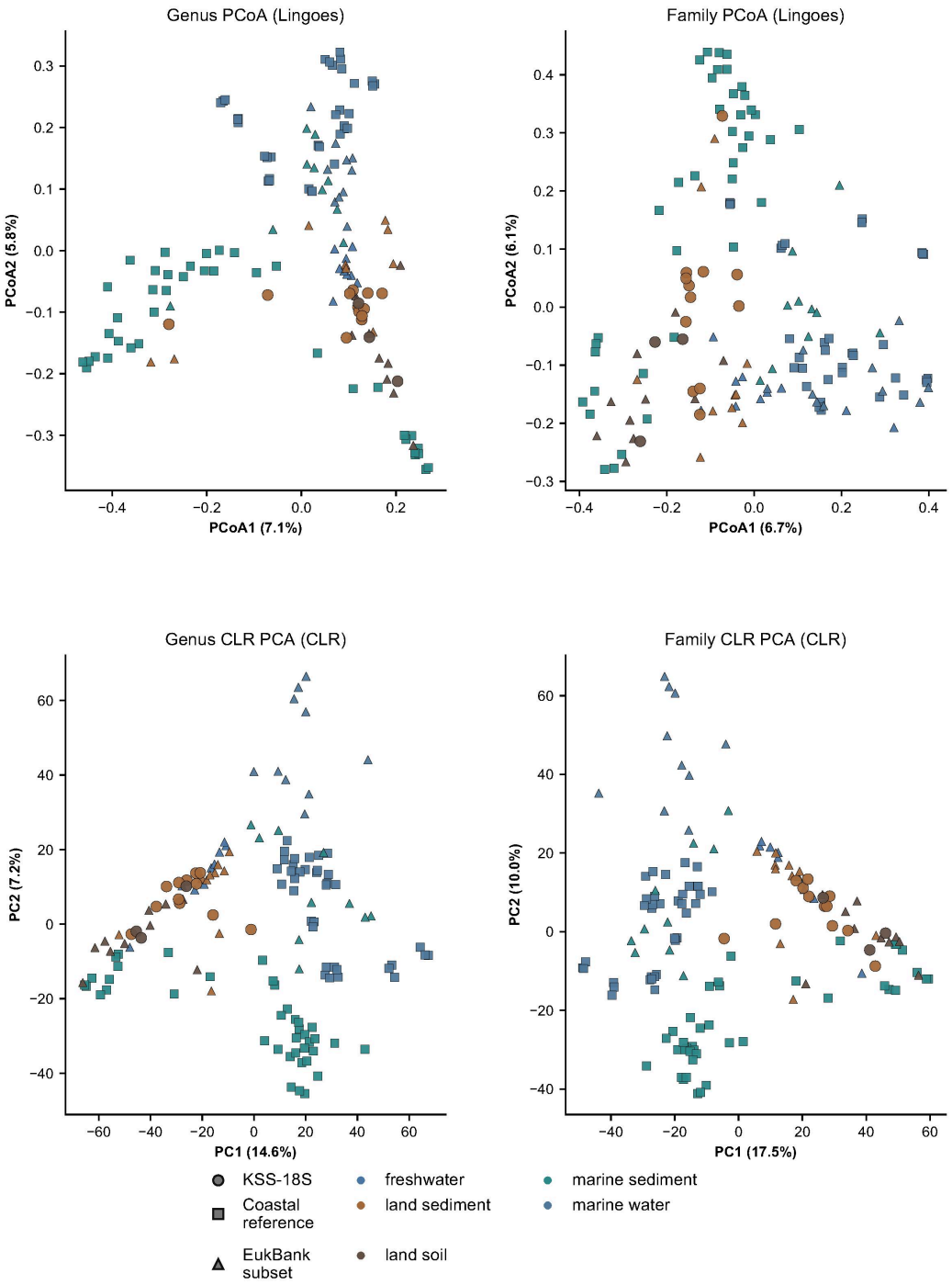


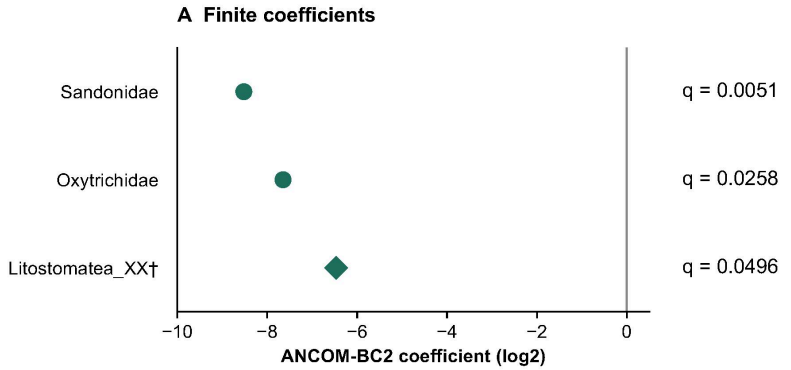
Figure 3

Comparative community structure

Genus and Family Bray–Curtis principal coordinates analysis (PCoA) with Lingoes correction (upper panels) and principal component analysis of centred log-ratio-transformed compositions (CLR-PCA; lower panels) for 140 libraries: 15 KSS-18S, 75 Coastal and 50 selected EukBank. Shape identifies dataset and colour identifies operational sample type. Axis labels report variation within the corresponding distance or log-ratio representation and are not directly comparable between the two methods. Compositions are normalized within assigned reads, retaining identifiable PR2 placeholders. Table S8 reports the tests and their permutation restrictions, including the finer-block sensitivity analysis, for which the restricted dispersion test was significant. Selection and assignment coverage are provided in Table S3. The finer-block result concerns the contrasts represented within blocks, not independent tests of all pairwise sample-type differences.

Figure 4. Exploratory Family-level habitat associations

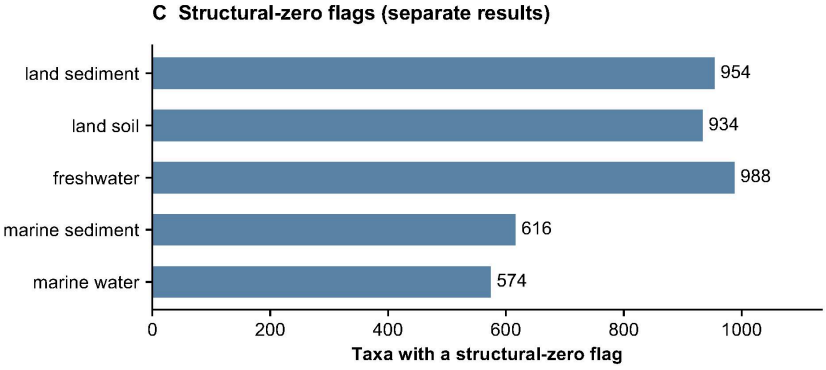
Clustered ANCOM-BC2; marine water relative to land sediment



† Singular random-effect fit; interpret with additional caution.
Negative coefficients indicate lower model-estimated abundance in marine water relative to land sediment.

B Full contrast inventory (finite-effect results)

Contrast vs land sediment	Finite coefficients	Robust candidates
land soil	64	0
freshwater	64	0
marine sediment	64	0
marine water	64	3



All four volcano panels are provided in Figure S26.
Robust means passing the recorded pseudo-count screen, not external validation.

Figure 4

Exploratory between-study Family associations

(A) The three retained pseudo-count-robust coefficients from analysis of compositions of microbiomes with bias correction 2 (ANCOM-BC2) for marine water relative to land sediment, with their recorded Benjamini–Hochberg q values. The model includes sample type and dataset with source-specific cluster

random intercepts. Points show bias-corrected log2 coefficients; no confidence intervals were constructed. The diamond and dagger flag the singular Litostomatea_XX random-effect fit. Oxytrichidae and Sandonidae have nonsingular fits but remain exploratory associations, not validated markers. No source study contained both marine-water and land-sediment libraries, so this is a between-study comparison. Source-cluster adjustment and pseudo-count robustness do not establish an independent habitat effect. (B) All four reference-coded contrasts retain 64 finite coefficients each; robust eligible candidates number zero for land soil, freshwater and marine sediment and three for marine water. (C) Structural-zero flags are counted separately and are not finite coefficients. The complete four-panel volcano display, including BH-significant non-robust and nonsignificant terms, is provided in Figure S26. There, the dashed q threshold is 0.05; nominal-only means Benjamini–Hochberg (BH) $q < 0.05$ but failure of the additional pseudo-count screen, not an unadjusted P value. Coefficients and q values are in Table S14 and candidate qualifications in Table S15.

Figure 5. Study-held-out ecological-context prediction

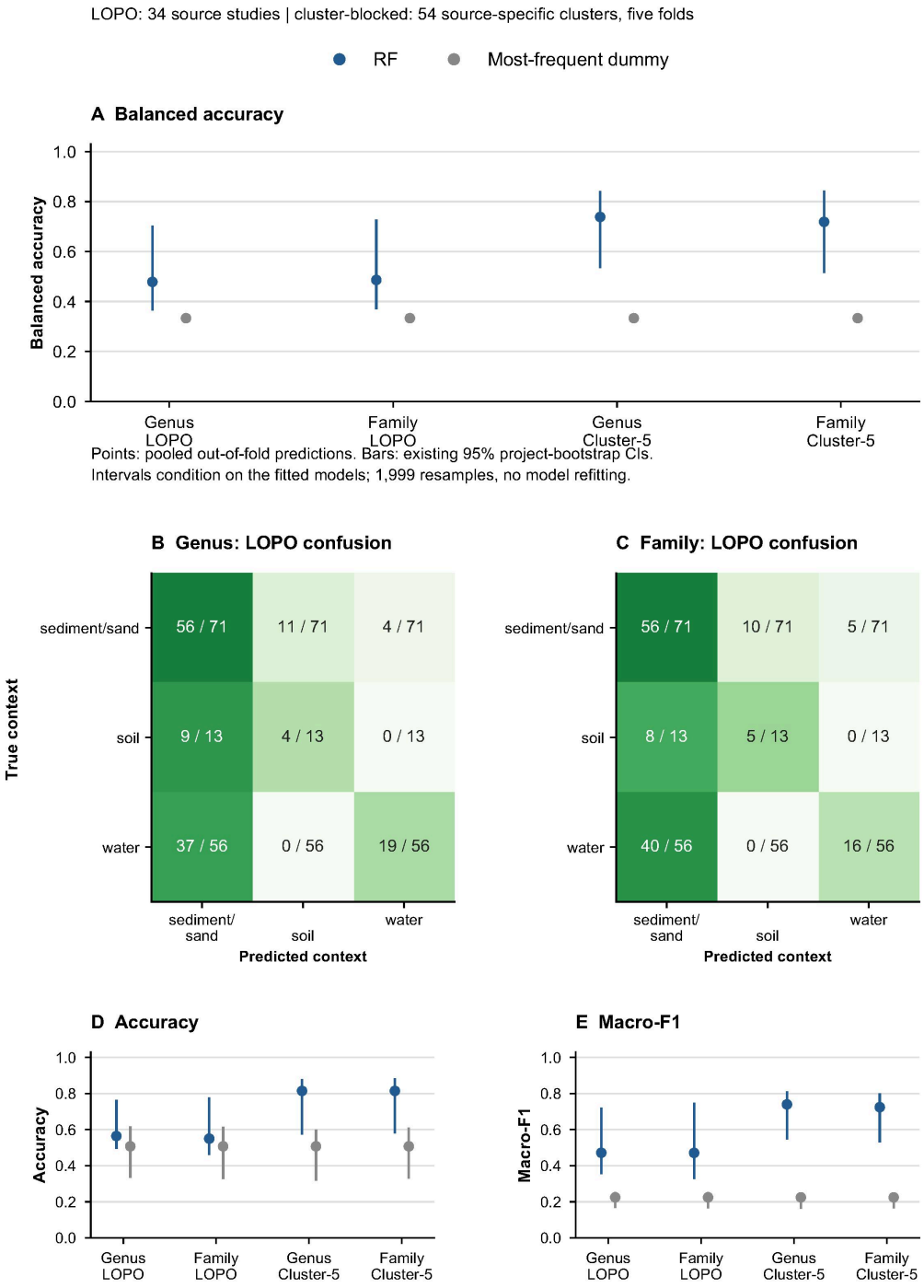


Figure 5

Ecological-context classification with source studies held out

(A) Pooled balanced accuracy, (D) accuracy and (E) supplementary macro-F1 for the matched target of 71 sediment/sand, 13 soil and 56 water libraries. Random Forest (RF) is compared with a most-frequent dummy classifier. Leave-one-project-out (LOPO) withholds each of 34 entire source studies; Cluster-5

allocates 54 source-specific clusters to five held-out folds and can retain other clusters from a test study in training. The 54 units comprise 15 KSS libraries, seven Coastal sites and 32 EukBank studies. Points are pooled out-of-fold estimates. Conditional 95% intervals use 1,999 source-study-block resamples of fixed predictions, without model refitting. (B,C) Genus and Family LOPO confusion matrices show counts relative to true-class totals. No paired significance test compared the validation designs. Balanced accuracy and confusion matrices are the primary discrimination summaries; macro-F1 is secondary. Complete predictions, equal-study summaries and settings are in Table S16 and SI Materials and Methods S5.

Supplementary Files

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

- Additionalfile1SupportingInformation.docx
- Additionalfile2SupplementaryTablesS1S18.xlsx
- Additionalfile3AdditionalDataTablesAD1AD19.xlsx
- Additionalfile4AdditionalDataFiguresAD1AD3.pdf