Microbial and Biochemical Surveys Along the Geographical Gradient of the Oregon Coast

by

Hengameh Habibi Soufi

A thesis accepted and approved in partial fulfillment of the

requirements for the degree of

Master of Science

in Biology

Thesis Committee:

Dr. Stilianos Louca, Chair

Dr. Peter Ralph, Member

Dr. Matthew Barber, Member

University of Oregon

Fall 2024

THESIS ABSTRACT

Hengameh Habibi Soufi

Master of Science in Biology

Title: Microbial and Biochemical Surveys Along the Geographical Gradient of the Oregon Coast

To understand the factors shaping microbial assembly and community composition in intertidal sediments along the Oregon coast, we collected sediment samples from 10 cores at depths of 160–190 cm, spanning over 300 km. Using metagenomic and 16S rRNA data, we constructed functional and taxonomic profiles, revealing redundancy in functional groups but variability in taxonomic composition. Regression models using chemical data measured in lab, geographical distances, and environmental variables from public datasets failed to significantly predict taxonomic composition. However, permutation null model tests suggested that biological interactions play a significant role in driving taxonomic variation in subsurface sediments, namely through taxa excluding each other via different mechanisms.

This thesis includes co-authored material submitted for publication in peer-reviewed journals.

TABLE OF CONTENTS

Chapter **Pages**

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

Table **Page**

LIST OF SCHEMES

Table **Page**

1. INTRODUCTION

Subsurface marine sediments form an extensive microbial habitat, covering about 70% of Earth's surface and playing a crucial role in global biogeochemical cycles (Hoshino et al., 2020; Schippers et al., 2012). These sediments host diverse microbial communities, including methanogens, sulfate-reducers, and other prokaryotes, which contribute significantly to nutrient cycling, organic carbon deposition, methane fluxes, and energy flow within marine ecosystems (Falkowski et al., 2008; Nealson, 1997; Schippers et al., 2012). The estimated abundance of microbial cells in these sediments is approximately 2.9×10^{29}, similar to microbial abundances in seawater and soil, accounting for roughly 0.6% of Earth's total living biomass (Kallmeyer et al., 2012). These microbes are essential for sustaining Earth's deep biosphere by influencing carbon and sulfur cycles, and they predominantly belong to uncultivated phylogenetic lineages that play key roles in organic matter degradation and elemental cycling (D'Hondt et al., 2004; Schippers et al., 2012; Parkes et al., 2014; Teske, 2006). This highlights the need for further research into the processes that support life in subsurface environments.

Despite the importance of these microbial communities, there is still limited understanding of their global distribution, diversity, and abundance, and the factors affecting microbial community assembly remain largely unexplored due to sampling challenges. Most previous research has focused on the vertical distribution of microbial taxa and genes along sediment columns (Edlund et al., 2008; Durbin & Teske, 2011; Molari et al., 2012; Vuillemin et al., 2018; Hoshino et al., 2020), demonstrating that redox conditions and thermodynamics are key factors in shaping these profiles (Canfield & Thamdrup, 2009; Wang et al., 2017). Other studies, which analyzed microbial composition across 40 globally distributed sampling sites at sediment depths ranging from 0.1 to 678 m, found that oxygen availability and organic carbon

concentration can partially predict the taxonomic composition and diversity of marine sediment communities (Hoshino et al., 2020). However, further exploration is needed to understand these relationships more comprehensively.

Another unresolved question is whether the metabolic functions of microbial communities in subsurface sediments are decoupled from their taxonomic composition within functional groups. Such decoupling can occur when mechanisms that constrain function, such as reaction stoichiometry, resource limitations, and physical transport bottlenecks, are independent of the mechanisms controlling which taxa perform each function (Louca, 2017; Louca et al., 2018; Louca et al., 2022). This decoupling would suggest that crucial ecosystem processes, such as nutrient cycling, may remain unaffected by changes in taxonomic composition (Fernández et al., 1999). While this phenomenon has been observed in other microbial systems, such as bioreactors, seawater, soil and host-associated microbiomes (Louca et al., 2018; Louca et al., 2020; Louca, 2021), it has not yet been confirmed in subsurface marine sediments.

Confirming this pattern in subsurface sediments would raise questions about how dispersal, abiotic environmental factors, and biological interactions determine which taxa occupy specific metabolic niches in different geographic locations, and at which spatial scales these factors become important. For instance, microbial dispersal rates in the subsurface are much slower than in surface environments (Louca et al., 2021; Anderson, 2021; Schippers et al., 2012; Parkes et al., 2014) which could cause less homogenization across space and potentially induce greater compositional differences between locations.

Understanding the complexities of microbial community assembly in subsurface marine sediments is essential for grasping the broader implications for global biogeochemical cycles and ecosystem stability. Despite considerable progress, significant knowledge gaps remain,

particularly concerning the factors influencing microbial community dynamics and functional roles in these environments. This study aims to address these gaps by focusing on two primary objectives. The first objective is to explore the factors affecting microbial assembly and taxonomic variability in subsurface sediments within intertidal zones along the Oregon coastline. This exploration will emphasize the influence of abiotic factors, the impact of dispersal using distance between sampling locations as a proxy, and the role of biological interactions in shaping community composition and function.

The second objective is to investigate the relationship between the taxonomic and functional composition of the microbial communities in these subsurface sediment samples. Understanding whether the functional roles are directly linked to the taxonomic identity of the microbes or if functional composition is independent from taxonomic diversity could be crucial for understanding ecosystem responses to environmental changes.

All chapters of this thesis were made possible with the supervision and contributions of co-author Dr. Stilianos Louca, and we plan on publishing the findings along with co-authors Robert E. Porch, Masha V. Korchagina, Joseph A. Abrams, Jared S. Schnider, Ben D. Carr, and Mark A. Williams.

2. DATA

The method for sampling sediment and pore water described in this chapter was primarily developed by me with the input from Dr. Stilianos Louca as well as other co-authors, Robert E. Porch, Masha V. Korchagina, Joseph A. Abrams, Jared S. Schnider, Ben D. Carr, and Mark A. Williams. All field trips were organized and carried out by me. All locations were investigated and chosen by me with input from other co-authors and Oregon State Park officials. All chemical measurements and molecular processing have been done by me. All computational analysis and visualizations have been done by Dr. Stilianos Louca. Writings for this chapter have been done by me and Dr. Stilianos Louca. All findings from this chapter are submitted for publication in a peer reviewed journal.

2.1 Methodology:

2.1.1 Sampling locations:

Samples were collected from marine coastal subsurface sediments at 10 different locations along the Oregon coast, covering a wide geographic range from north to south to capture environmental variability (Fig. 1, Supplemental Table S1). All samples were taken from the intertidal zone, at a consistent depth range of 1.6–1.9 m. This depth was selected as it represented the maximum penetration achievable across all locations while minimizing sampling complications.

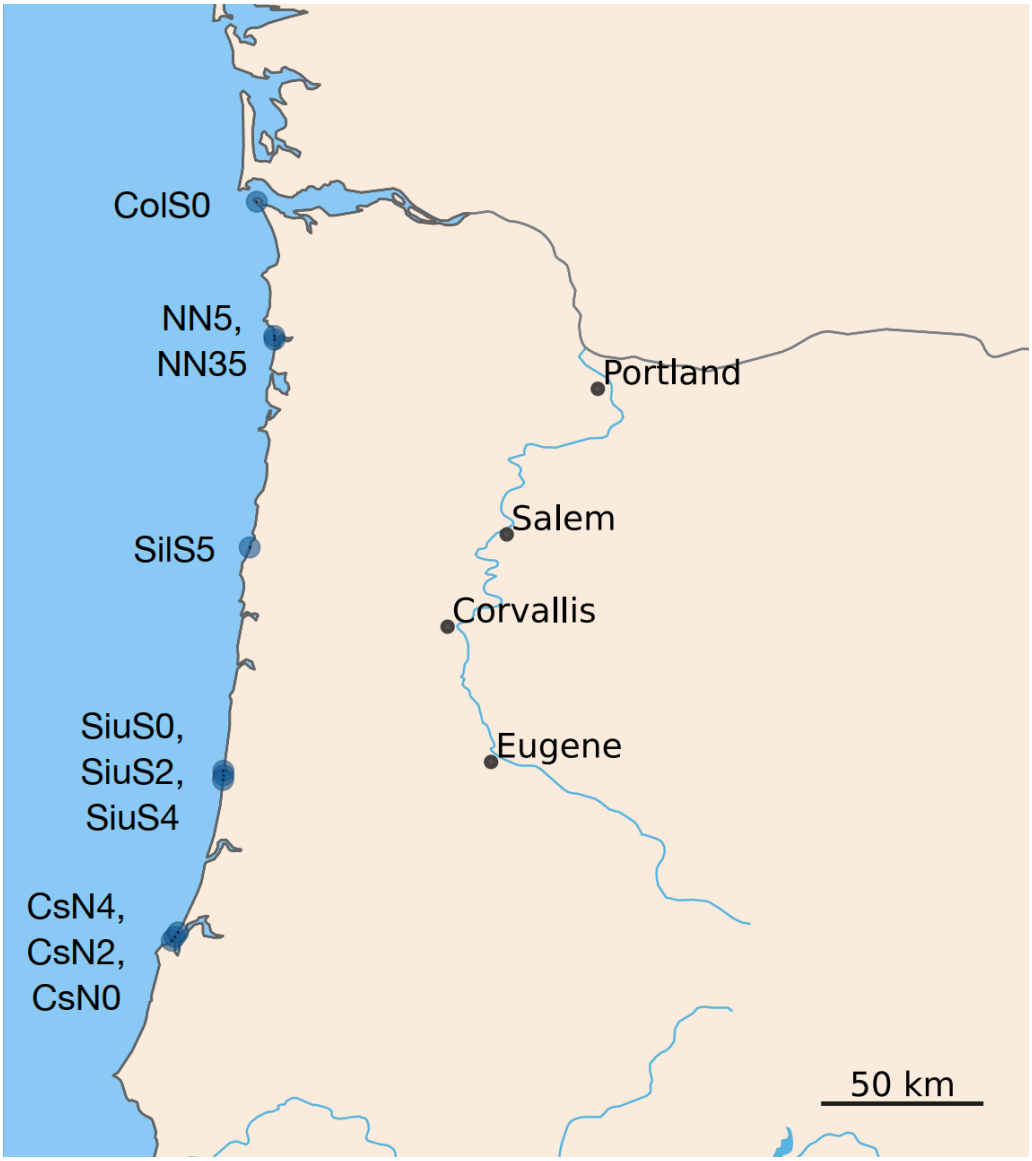

Figure1. Sampling Locations. Map of Oregon coastline roughly depicting the 10 sampling locations.

The locations were chosen based on accessibility, as transporting heavy equipment to and from the sampling sites required logistical feasibility. Additionally, the selected locations varied in their proximity to river deltas to include different environmental conditions, such as changes in salinity, nutrient concentrations, and temperature, which are expected to influence microbial community composition. The sampling sites were also separated by at least 1.5 km and spanned over an area of more than 300 km along the Oregon coast to ensure sufficient spatial coverage.

2.1.2 Sampling procedure:

A galvanized steel pipe (6.045 cm diameter) was pushed into the ground using a gasoline-powered post driver, sealed at the top using a Gripper® expanding pipe plug, and subsequently pulled out using a farm jack affixed to the pipe with a Morris coupling. This sampling depth was chosen as the maximum depth that could be practically reached with an 8-foot steel pipe, which in turn was the practical size limit given the transportation means permitted at some locations, the equipment needed for inserting and extracting the pipe, and ultimately our overall budget. A standard sampling procedure is depicted in Figure 2.

Figure 2. **Standard sampling procedure.** Photo taken by me south of Columbia River delta, OR. Overview of sampling locations, dates can be seen in Table S 1.

While we have no concrete reason to expect our overall conclusions to only be valid for this sampling depth, the generality of our findings can only be fully confirmed by future studies examining alternative depths. Material for DNA extraction was collected in 50 ml centrifuge tubes after breaking the seal created by the gripper plug and sliding out 30 cm of the deeper end of sediment core. Water for chemical analyses was collected from the same cores by centrifuging the collected sediment and transferring the supernatant water to second tubes for storage. All samples were stored on dry ice in the field and subsequently at -80°C in the laboratory until further processing.

2.1.3 Genomic analysis:

DNA was extracted from the collected sediments using the Qiagen™ DNeasy Power Biofilm kit following the manufacturer's protocol. To reduce spurious variance in microbial community compositions merely due to microheterogeneities and due to the coarseness of the core's depth estimates, 3 extractions were performed from each core from nearby layers and subsequently pooled, roughly equally spaced within the depth range 1.6–1.9 m. Shotgun metagenomic and 16S rRNA gene amplicon sequencing was performed for each of the 10 samples by the Integrated Microbiome Resource (IMR) in Dalhousie, Canada. Specifically, metagenomic libraries were prepared using the Illumina Nextera Flex kit and sequenced using a NextSeq2000 (2×150 bp paired ends). 16S rRNA gene amplicon fragments (V4-V5 region) were PCR-amplified using the Phusion Plus polymerase and "universal" bacterial + archaeal primers (515FB = GTGYCAGCMGCCGC-GGTAA, 926R =CCGYCAATTYMTTTRAGTTT; (Parada et al., 2016; Walters et al., 2015)) and sequenced on a MiSeq (2×300 bp paired ends).

Analysis of 16S rRNA gene amplicons

On average 11846 16S rRNA gene read pairs were obtained for each sample. Reads were quality-filtered, and amplicon sequence variants (ASVs) were inferred and chimera-filtered using the R package dada2 v1.28.0 (Callahan et al., 2016), as follows. Reads were quality-filtered using the dada2 function filterAndTrim, with options "truncLen = (250,200), maxEE=(1,1), truncQ=(0,0), trimLeft=(6,6), minLen = (100,100), maxLen = (100000,100000)", retaining on average 9327 read pairs per sample. Error model calibration for ASV inference was performed jointly for all samples but separately for forward and reverse reads. Calibration was performed using the dada2 function learnErrors with options "nbases=1e8, randomize=TRUE, MAX_CONSIST=10, errorEstimationFunction=loessErrfun". Reads were dereplicated using the dada2 function derepFastq. ASVs were inferred from the dereplicated sequences separately for forward and reverse reads, using the dada2 function dada (options "pool=TRUE, selfConsist=FALSE") and the previously calibrated error models. ASVs from forward and reverse reads were merged using the dada2 function mergePairs with options "minOverlap=12, maxMismatch=0, trimOverhang=TRUE". Merged ASVs were chimera-filtered using the dada2 function removeBimeraDenovo (option method = 'consensus'). This yielded an ASV table of 4380 chimera-filtered ASVs, accounting for 54116 reads across 10 samples.

ASVs were taxonomically classified based on a comparison to the SILVA database v138.1 (Glöckner et al., 2017), using a consensus approach described previously (Soufi et al., 2024). In total, 1 ASV was identified as chloroplast, no ASVs were identified as mitochondria, and 64 ASVs could not be classified at any taxonomic level; these ASVs were omitted from subsequent analyses. This left us with 4315 prokaryotic ASVs, accounting for 53477 reads across all samples. To remove species-level redundancies in ASVs, we also clustered ASVs into

operational taxonomic units (OTUs) de-novo at 99% similarity, a commonly accepted operational threshold for species delineation (Kim et al., 2014). Clustering was done using vsearch-cluster_fast with options "-iddef 2-strand plus", which yielded 1526 prokaryotic OTUs. Taxonomic identities of OTUs were inherited from their representative (centroid) ASVs.

Pairwise dissimilarities between samples in terms of microbial taxonomic composition (ASV and OTU levels) were computed using 3 different metrics, all of which accounted for ASV/OTU abundances: Bray-Curtis, Hellinger, and Jaccard (Hibbing et al., 2010). These dissimilarity matrices were used for Mantel tests and regression analysis, described below.

Analysis of metagenomes

On average, 4,788,369 metagenomic read pairs (~1.2 Gbp) were obtained per sample. Adapters were trimmed from reads using the tool skewer v0.2.2 (Jiang et al., 2014). Reads were then quality-filtered using vsearch v2.22.1 (Rognes et al., 2016) with options "-fastq_ascii 33 -fastq_maxee 0.2 -fastq_truncee 0.2 -fastq_qmax 64 -fastq_maxee_rate 0.002 -fastq_stripleft 0 -fastq_trunclen_keep 10000", retaining on average 3,164,169 high-quality read pairs per sample. High-quality read pairs were merged using flash v1.2.11 (Magoc & Salzberg, 2011), with options "-min-overlap 12 --max-overlap 300 --phred-offset 33 --max-mismatch-density 0.01 -allow-outies". Any forward reads from unmerged high-quality pairs were then combined with the set of merged sequences. Sequences (merged read pairs and unmerged forward reads) shorter than 100 bp were subsequently omitted from all downstream analyses. Sequences from all 10 samples were co-assembled into longer contiguous sequences (contigs) using megahit v1.2.9 (Li et al., 2015) with option "-min-contig-len 500". A total of 133,241 contigs were generated, with an average length of 821 bp, a maximum length of 44,434 bp, and an N50 of 780.

Gene-centric functional profiles were generated from assembled contigs as described previously (Soufi et al., 2024). Here we thus only provide a brief summary. Contig coverages were computed for each sample by mapping the non-assembled sequences to the contigs, then counting the number of sequences mapped to each contig with a MAPQ score ≥30 and dividing that number by the contig length. Contig coverages were then normalized in each sample to sum 1, thus yielding contig "proportions". Protein-coding genes (PCGs) were predicted in the contigs using prodigal v2.6.3 with option "-p meta" and otherwise default options (Hyatt et al., 2010). PCGs were then either mapped to KEGG gene orthologs (KOs) in the KOfam HMM database r105 (Aramaki et al., 2019), or mapped to the AsgeneDB amino-acid sequence database of arsenic-metabolism-related genes (Song et al., 2022), or mapped to a custom set of perchlorate reduction genes (pcrABC). Only hits with an E-value below 10^{-10} were considered. Proportions of PCGs were computed in each sample by first associating with each PCG the proportion of its host contig and then normalizing those values in each sample to sum 1; in other words, PCG proportions express the relative abundance of each PCG in a sample compared to all predicted PCGs. The proportion of a given gene in a given sample was estimated by summing the proportions of all proteins mapped to the specific gene.

To obtain profiles of the functional potential of each sampled microbial community, genes were assigned to custom functional groups described previously (Soufi et al., 2024, Table S3 therein). KOs were also grouped into standard KEGG categories, at hierarchical levels A, B, and C. In addition, KOs were grouped according to their Enzyme Commission (EC) numbers (International Union of Biochemistry and Molecular Biology, 1992), which correspond to distinct enzymatic functions and provide the highest meaningful resolution of functions. Note that the individual KOs represent level D in the KEGG hierarchy; since KOs are defined based

on orthology and not based on function, level D profiles are not strictly speaking functional profiles and are thus not considered here. The proportion of each functional group (or KEGG category or EC) in each sample was computed as the sum of proportions of all associated genes. The following KEGG categories were omitted, as they are not actually defined based on function: "brite hierarchies", "enzymes with ec numbers", "not included in pathway or brite", "poorly characterized", "general function prediction only", "others", "unclassified viral proteins", "function unknown".

2.1.4 Environmental variables:

In-situ measurements

Pore water collected from the samples was filtered through a 0.22 µm syringe filter in the lab after defrosting. The filtered pore water was then used to measure various in-situ environmental properties. Salinity was determined using a refractometer, while sulfate concentrations (mg/L) were measured using a Hach® DR1900 spectrophotometer and the Hach® TNT 865 sulfate kit. Nitrate measurements were attempted using the Hach® TNT 836 kit, but these were unsuccessful due to very low concentrations.

The concentrations of boron, barium, calcium, potassium, magnesium (I and II), manganese, sodium, sulfur, silicon, and strontium were quantified using Inductively Coupled Plasma Optical Emission Spectroscopy (ICP-OES) at the Keck Laboratory, Oregon State University. Due to the variability in element concentrations, samples were diluted up to 100-fold before measurement.

Public data

Annual average values of regional environmental variables used for model building were determined for each sampling location based on publicly available datasets, accessed on May 21,

2024. Monthly average ocean surface temperatures, surface chlorophyll concentrations, and solar insolations were downloaded from the NASA Earth Observations gridded database, with a spatial grid resolution of 0.25° (dataset IDs MYD28M, MY1DMM_CHLORA, and CERES_INSOL_M, respectively), and subsequently averaged over the 12 months preceding our sample collections (October 2022 – September 2023). Monthly multi-year average ocean surface nitrate, phosphate, and silicate concentrations were downloaded from the World Ocean Atlas database release 2023 (Reagan et al., 2024), hosted by the US National Centers for Environmental Information, accession 0270533, with a spatial grid resolution of 1°, and subsequently averaged over all 12 months. Monthly multi-year average ocean surface salinities were downloaded from the World Ocean Atlas database release 2023, with a spatial grid resolution of 0.25°, and subsequently averaged over all 12 months. Gridded data were evaluated at sample locations via bilinear interpolation. If a sample was located inside a grid cell where values were missing on some or all cell corners, interpolation was done using a triangulation of the grid points with non-missing values.

2.1.5 Statistical analysis:

Regression analysis and Mantel tests

To examine the potential role of dispersal on microbial community structure, we performed Mantel rank correlation tests (Legendre & Legendre, 1998), comparing pairwise dissimilarities in community composition to pairwise geographic distances. We considered three of the most common dissimilarity metrics, Jaccard, Bray-Curtis, and Hellinger (Legendre & Legendre, 1998; Legendre & Gallagher, 2001; Chao et al., 2005), calculated at the level of ASVs as well as OTUs. The one-sided statistical significance of Spearman rank correlations was estimated through 1000 random permutations of the dissimilarity matrix's rows and columns,

each time permuting rows and columns in the same manner, as is standard procedure in Mantel tests. An overview of dissimilarities and geographic distances is shown in Supplemental Fig. S1.

To examine the potential ability of environmental variables to explain taxonomic composition differences between samples, we attempted to build linear regression models, whose response variables were pairwise dissimilarities (Bray-Curtis, Jaccard, or Hellinger) in taxonomic composition (at ASV level, or OTU level, etc.) and whose potential predictor variables were pairwise geographic distances as well as pairwise absolute differences in the 20 physical-chemical variables mentioned earlier (regional environmental variables, ICP-OES measurements, in-situ salinity). Hence, each sample pair constituted a single datapoint for the model. Linear coefficients were fitted using least squares, and predictors were added one at a time using a stepwise approach whereby a predictor was only included if its linear coefficient was statistically significantly different from zero ($P < 0.05$). This statistical significance was assessed through simultaneous permutations of the response matrix's rows and columns, similar to the permutation null models commonly deployed in Mantel tests, following Legendre, Lapointe, & Casgrain (1994).

Analysis of taxon co-distributions

To examine potential interdependencies between taxon distributions across samples, we performed two alternative null hypothesis tests separately for each taxonomic level (ASVs, OTUs, genera, etc.). Both tests are commonly used in ecology to detect non-neutral patterns in the joint distributions of multiple species, for example resulting from competitive exclusion or mutualisms (Gotelli, 2000). Each test defines a test statistic that conceptually corresponds to a notion of taxon overlaps, or a correlation in the distribution of taxa, as well as a null model for generating random data for computing statistical significances. In the first test, we considered a

summary statistic computed based on the presences/absences of taxa in each sample, henceforth referred to as *checkerboard cooccurrence* (CC) score:

Equation 1

$$\mathrm{CC} = 1 - \frac{\sum_{i=1}^{M}\sum_{j=1}^{i-1}(N_i - N_{ij})(N_j - N_{ji})}{\sum_{i=1}^{M}\sum_{j=1}^{i-1}(N_i - Np_ip_j)(N_j - Np_ip_j)}$$

where *M* is the total number of considered taxa, *N* is the total number of samples, *N*i is the number of samples containing the *i*'th taxon, N_{ij} is the number of samples containing both taxa *i* and *j*, and p_i := *N*i/*N*. Hence, for fixed *N1, N2, ...*, N_M, the CC-score becomes larger if taxa co-occur more frequently (i.e. N_{ij} are larger). Note that this summary statistic is closely related to the "C score" described by Gotelli (2000), with two differences: The CC score is normalized differently such that its scale remains roughly constant as the number of samples increases, and it is reversed such that a greater value corresponds to a greater overlap in taxon distributions. To assess whether an observed CC score was probably due to chance (i.e., if taxa occur independently of each other), we compared it to the CC score distribution of 1000 presence-absence matrices randomly generated under a null model. As null, we used the "SIM9" permutation model described by Gotelli (2000), also known as the "fixed-fixed" model because it preserves the number of taxa per sample and the number of samples per taxon (Kallio, 2016; Louca et al., 2016). If random CC scores generated by the null model are mostly above the observed CC score, this would mean that taxa tend to exclude each other more often than expected by chance (i.e., taxa are segregated). To account for multiple hypothesis tests (i.e., one

for each taxonomic level), we also considered a Bonferroni-adjusted significance threshold of $\alpha\tilde{}$ = $\alpha/n = 0.05/7 = 0.0071$. An overview of results is shown in Supplemental Table S5.

In the second test, we considered a summary statistic based on the relative abundances of taxa in each sample, known as *generalized Morisita similarity index* (Chao et al., 2008) and henceforth referred to as "MA-score" (Ulrich & Gotelli, 2010):

Equation 2

$$\mathrm{MA} = \frac{\sum_{i=1}^{M}\left[\left(\sum_{j=1}^{N} p_{ij}\right)^2 - \sum_{j=1}^{S} p_{ij}^2\right]}{(N-1)\sum_{i=1}^{M}\sum_{j=1}^{N} p_{ij}^2},$$

where p_{ij} is the relative abundance of taxon *i* in sample *j*. Hence, a lower MA score indicates a lower similarity between samples in terms of taxon abundances and thus a potential segregation between taxa. As null, we considered the "IT" model suggested by Ulrich et al. (Ulrich & Gotelli, 2010), which assigns sequences to matrix cells proportional to the total number of sequences in each sample and proportional to the total number of sequences assigned to each taxon across samples, until the total number of sequences per sample and per taxon is reached (Louca et al., 2016). We used 1000 abundance matrices randomly generated by this model to assess the statistical significance of MA scores. An overview of results is shown in Supplemental Table S6.

2.2 Results:

2.2.1 Microbial community composition

To elucidate the functional structure of the surveyed microbial communities, we used gene-centric metagenomic sequencing. After assembling reads into contiguous sequences (contigs), predicting and annotating protein-coding genes in the contigs, we determined gene proportions based on reads mapped to contigs. We then classified genes into groups at each functional classification level of the KEGG hierarchy (A to C), with A being the coarsest classification level (e.g., metabolism vs cellular processes), B being a finer level (e.g., energy metabolism vs lipid metabolism) and C being an even finer level (e.g., oxidative phosphorylation vs photosynthesis) corresponding to individual KEGG pathway maps (Kanehisa et al., 2004, 2021). In addition, we grouped genes according to catalyzed reactions based on enzyme commission numbers (ECs; International Union of Biochemistry and Molecular Biology, 1992), which represent the finest available and meaningful functional classification. At KEGG level A, profiles were dominated by genes involved in metabolism, followed by genes involved in genetic information processing, and genes involved in environmental information processing (Fig. 3A). At KEGG level B, profiles were dominated by genes involved in genetic information processing, genes involved in signaling and cellular processes, and genes involved in metabolism (Fig. 3B). At each considered KEGG level, all samples had nearly identical functional structure in terms of the proportions of the various gene groups (Fig. 3). A similar observation was made for ECs and for our custom gene groups (Fig. 3D). This suggests that the metabolic pathways and ecological functions of the local microbial communities are of similar importance across samples, and their proportions strongly constrained by similar redox conditions and stoichiometric balances (Canfield & Thamdrup, 2009; Louca et al., 2018). This consistency of functional structure is

particularly remarkable given that these samples were obtained along a transect that span over 300 km. In fact, this transect intersects the deltas of multiple major rivers such as Columbia, Siuslaw, Nehalem, and Coos, originating in distinct geographic regions and with discharge rates spanning 3 orders of magnitude (Supplemental Table S2).

To elucidate the taxonomic composition of microbial communities, we used 16S rRNA gene amplicon sequencing, with reads either resolved at the level of individual amplicon sequence variants (ASVs; Callahan et al., 2016), or clustered into operational taxonomic units (OTUs, 99% similarity; Kim et al., 2014; Edgar, 2018) or grouped into higher-order taxa. At the genus level, the most abundant taxa were Subgroup 10 in the family Thermoanaerobaculaceae, followed by Woeseia, Blastopirellula, and Rhodopirellula (Fig. 4B). At the class level, microbial communities in all samples were dominated by Planctomycetes, Gammaproteobacteria, and Thermoanaerobaculia (Fig. 4C). The bulk of the communities, i.e., most of the reads, belonged to taxa whose proportions exhibited high variability across samples. Among the abundant taxa, only a small number had relatively stable proportions across samples. This general taxonomic variability was particularly strong at higher taxonomic resolutions, such as ASV or OTU level. Indeed, nearly all reads were mapped to ASVs and OTUs that exhibited strong fluctuations in their proportions across samples (Fig. 4A and Supplemental Fig. S2). This observation contrasts the much more stable functional structure across samples discussed earlier. This suggests that while the proportions of various functional groups are similar across all locations, the specific taxa encoding each function are highly variable. As a case in point, the average number of OTUs detected in each sample was 806.2, while the average number of OTUs shared by any two randomly chosen samples was only 478.6 (i.e., down by 40.6%), and the average number of OTUs found in all 10 samples was only 51 (Fig. 3D). This pattern is even stronger at the ASV

level: While each sample exhibited on average 1726 ASVs, the number of ASVs shared by any two samples was 821 (i.e., down by 52%), and the number of ASVs shared by all 10 samples was only 11 (Fig. 4D). This means that not only do the proportions of taxa change across samples, but many taxa also found in one sample can be absent (or at least below detection limit) in another sample. Such a decoupling between functional and taxonomic composition in microbial communities has been reported previously in other environments, notably in bioreactors (Fernández et al., 1999), human guts (Turnbaugh et al., 2009), in green macroalga (Burke et al., 2011), and bromeliad plants (Louca et al., 2016). Explanations previously proposed for this pattern generally invoke functional redundancy, i.e., the existence of multiple taxa capable of similar metabolic functions (Louca et al., 2018), combined with specific mechanisms promoting alternative taxa in any given functional group, such as phage-host dynamics (Louca et al., 2017), transport-limited metabolic activity (Louca et al., 2022), and antibiotic warfare between species (Czárán et al., 2002; Hibbing et al., 2010).

2.2.2 Role geographic distance

To assess whether distance-dependent dispersal limitation could explain the observed variation of taxonomic composition, we computed pairwise dissimilarities between samples and performed Mantel tests of Spearman rank correlations between dissimilarities and geographic distances (Legendre & Legendre, 1998), separately for each considered taxonomic level (ASV, OTU, ..., phylum). Dissimilarity metrics that we considered were abundance-based Bray-Curtis, Hellinger, and Jaccard (Legendre & Legendre, 1998; Legendre & Gallagher, 2001; Chao et al., 2005), which are commonly used in ecology. In contrast to simple correlation tests, Mantel tests are better suited for assessing the significance of correlations between distance matrices, as they

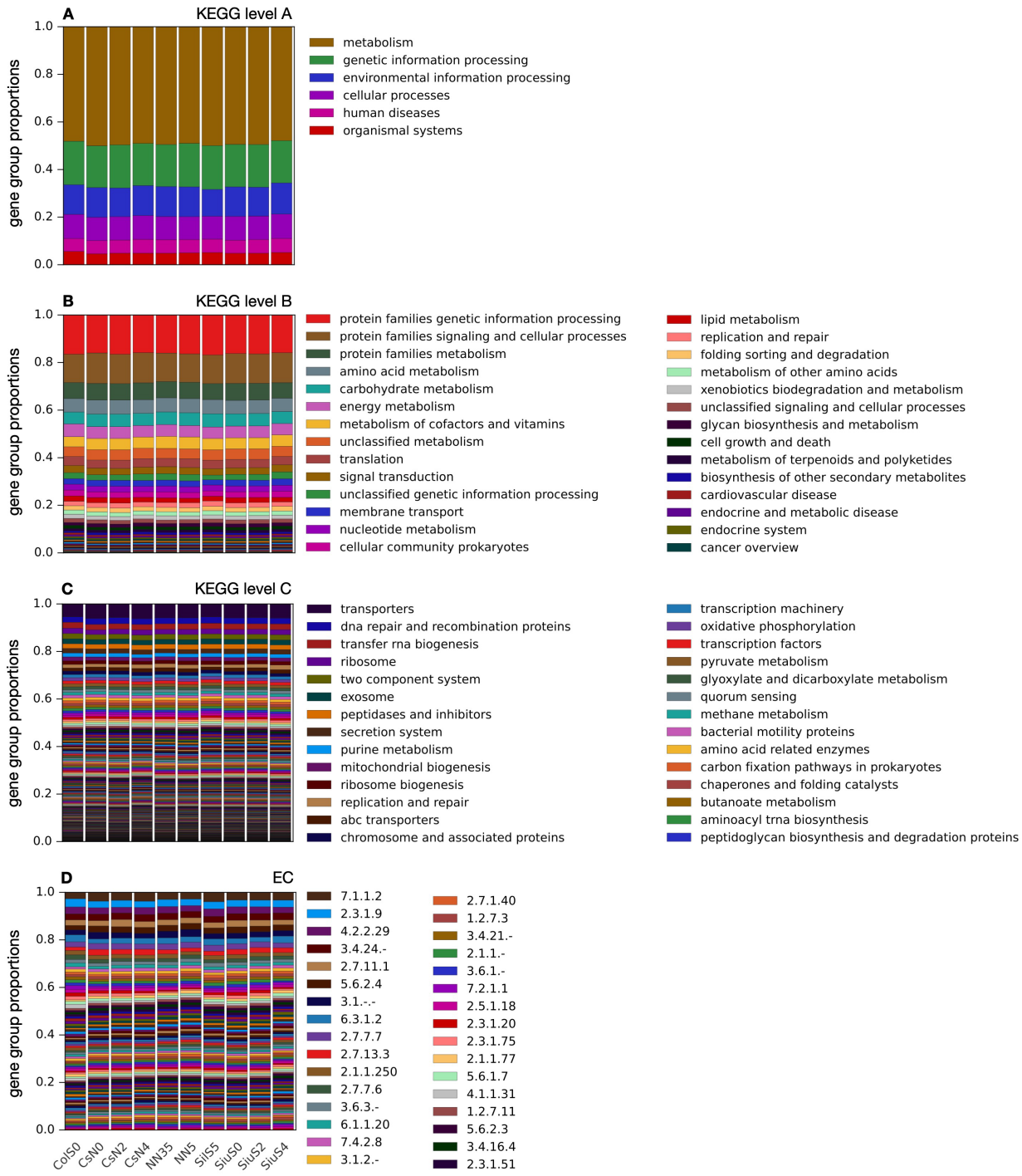

Figure 3 Gene group profiles (all KEGG levels). Estimated proportions (relative abundances) of genes associated with various standardized KEGG categories, at hierarchical levels A, B and C, as well as various Enzyme Commission (EC) numbers. Proportions are based on the average number of metagenomic reads mapped per protein base pair. Proportions are normalized in each sample such that their sum over all gene groups is 1. Only the most abundant gene groups are listed in the legends for simplicity. In (D) only the 100 most abundant ECs are shown for visibility.

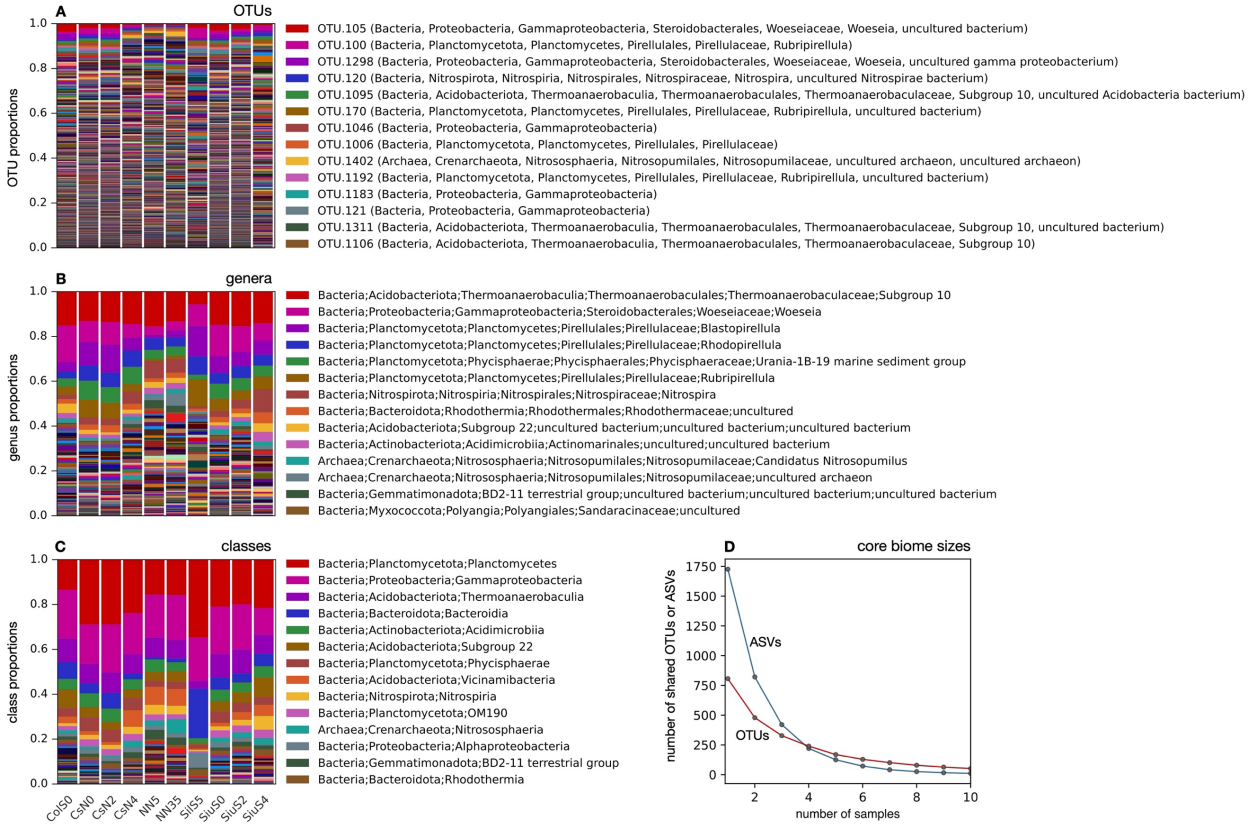

Figure 4 Taxonomic Compositions. (A) Proportions (relative abundances) of various prokaryotic Operational Taxonomic Units (OTUs, clustered at 99% similarity), based on 16S rRNA gene amplicon read counts. OTUs are sorted from top to bottom in decreasing average proportion. Only the top few OTUs are listed in the legend for readability. Estimated taxonomic identities of OTUs are written in parentheses. (B, C) Similar to A, but showing proportions of genera and classes, respectively. For a similar plot of amplicon sequence variant proportions see Supplemental Fig. S2. (D) Core biome sizes as a function of the number of samples. For any given number of samples n, the curves show the average number of ASVs or OTUs shared by n randomly chosen samples.

account for interdependencies between matrix entries stemming from the intrinsic data structure. In nearly all of the 21 tests (7 taxonomic levels × 3 dissimilarity metrics), correlations between dissimilarities and geographic distances were non-significant (details in Supplemental Table S3). Significant correlations were only observed at the phylum level using Jaccard and Hellinger dissimilarities, and at the genus level using Jaccard dissimilarity (correlations 0.34, 0.28, and 0.38, respectively, one-sided $P < 0.05$). In fact, if the significance threshold is adjusted to

account for multiple hypothesis tests using a Bonferroni correction ($\alpha = 0.05/21 = 0.0024$), none of the correlations remain significant. Thus, the influence of geographic distance on taxonomic differences at these spatial scales is not strong enough to be robustly detectable in our dataset. We stress, however, that this does not rule out the existence of dispersal limitation between sampling points. In fact, it is likely that dispersal between sampling points is severely limited over the time scales at which microbial communities typically change (Louca, 2021), and that correlations between distance and taxonomic composition would be more intense over much shorter distances than those examined here (Legendre & Gallagher, 2001).

2.2.3 The role of environmental conditions

To examine the role of environmental conditions as potential drivers of taxonomic variation across samples, we attempted to build linear regression models whose response variables were pairwise dissimilarities of taxon proportions. A separate model was built at each taxonomic level (ASV, OTU, ..., phylum) and for each considered dissimilarity metric (Bray-Curtis, Hellinger, or Jaccard). As possible predictor variables, we considered pairwise absolute differences in various environmental variables as well as pairwise geographic distances. Environmental variables included annual-average regional oceanographic variables from public gridded databases, such as surface temperature and surface salinity, surface concentrations of chlorophyll, nitrate, phosphate, and silicate, as well as concentrations of various elements (B, Ba, Ca, K, Mn, Na, S, Si, Sr), salinity, and sulfate concentrations that we measured directly in the collected pore waters (Fig. 5). Note that short-term (e.g., daily or weekly) fluctuations in surface water conditions are unlikely to impact microbial communities in the sampled subsurface layers, due to the slow transport of heat and dissolved substances across sediment columns (Caldwell, 1973; Iversen & Jørgensen, 1993; Louca et al., 2022) and the fact that microbial cell turnover

rates in marine subsurface sediments are generally slow (Thorn & Ventullo, 1988; Jørgensen, 2011; Hoehler & Jørgensen, 2013; Anderson, 2021). Predictors were selected one-by-one in a stepwise manner, keeping any predictors, whose coefficients were statistically significant ($P < 0.05$).

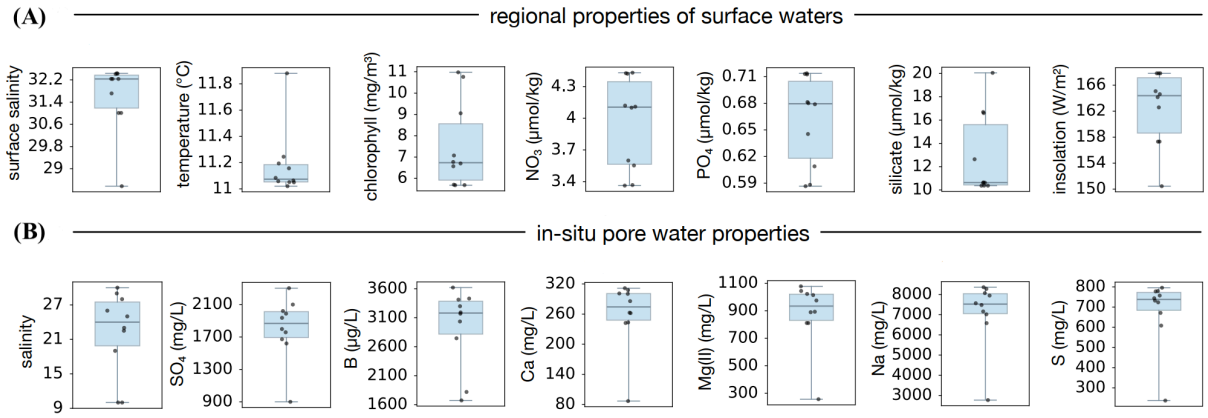

Figure 5 Regional and in-situ environmental variables. Scatter point represents one sample, boxes span interquartile, whiskers show the full data range and horizontal line segments show medians. **(A)** Annual-average regional environmental variables of surface waters, obtained from public gridded datasets and interpolated onto the sampling locations. **(B)** Salinity and sulfate concentrations measured in the pore waters collected from the cores and Concentrations of major elements measured in the pore waters collected from the cores.

We found that in nearly all cases, i.e., for most taxonomic levels and dissimilarity metrics, none of the considered variables were chosen as model predictors, that is, the majority of coefficients were non-significant (details in Supplemental Table S4). The only exceptions were models that predicted Hellinger dissimilarities at the phylum, class, family, or genus level, where boron concentration was selected as the sole predictor, achieving a fraction of explained variance during cross-validation (R^2) between 0.22 and 0.27. The selection of boron as a predictor in some cases appears surprising to us, and we are not aware of an obvious mechanism by which boron would influence microbial communities more strongly than other examined factors. We speculate that perhaps boron merely correlates with — and thus acts as a proxy for

— other non-measured environmental factors impacting or impacted by microbial communities. For example, boron can be strongly enriched in certain organic compounds in sediments (Williams et al., 2001) and is known to interact and bind with clay minerals (Kabay et al., 2015). If the significance threshold were to be adjusted for the multiple hypothesis tests across taxonomic levels, metrics, and candidate predictors ($\alpha = 0.05/(3 \cdot 7 \cdot 21) = 0.00011$), then neither boron nor any other predictor is selected. Thus, much of the taxonomic variation across samples was generally poorly explained by the considered environmental variables, suggesting that this variation is driven mostly by other factors. This finding reflects similar observations in surveys of other environments when functional composition was either constant or separately accounted for, for example in the foliage of bromeliads (Louca et al., 2016), in ocean waters (Louca et al., 2016), and in soils (Nelson et al., 2016).

2.2.4 Null model tests of community assembly

To examine the potential role of biological interactions, such as antagonism or mutualism, in the taxonomic composition of microbial communities, we performed two alternative null model tests commonly encountered in ecology (Connor & Simberloff, 1979; Hausdorf & Hennig, 2007; Louca et al., 2016). One test is based on taxon occurrence (i.e., presence/absence) patterns (Gotelli, 2000) and the other test is based on relative abundance patterns (Ulrich & Gotelli, 2010). Both tests define a summary statistic that measures the degree to which taxa overlap in their distributions across samples, and a null model from which random composition data can be generated for computing an expectation and statistical significance of the summary statistic. The two summary statistics are henceforth referred to as "CC score" and "MA score," respectively; precise mathematical definitions and details on the null models are given in the methods. A significantly high CC or MA score means that taxa tend to co-occur

more frequently than expected by chance, potentially due to positive interactions, while a significantly low CC or MA score means that taxa tend to segregate, potentially due to negative interactions.

Detailed results at each taxonomic level for the two tests are given in Supplemental Tables S5 and S6. We found that all CC scores and all MA scores were smaller than expected under the null model, regardless of taxonomic level. All MA scores were statistically significantly low ($P < 0.05$), even when adjusting the significance threshold for multiple hypothesis tests using a Bonferroni correction ($P < 0.0071$). Similarly, nearly all CC scores (except at phylum and class level) were statistically significantly low even after Bonferroni correction ($P < 0.0071$). This strongly suggests that taxa tend to segregate in their distributions. This result is particularly remarkable in view of the fact that the null model considered for the CC scores, known as the "fixed-fixed" model, is generally regarded as conservative, i.e., frequently accepting the null hypothesis (Gotelli, 2000). These results thus suggest that community assembly is strongly influenced by mutual exclusions between taxa, likely due to biological interactions. Similar null model analyses of microbial communities in bromeliad foliages also revealed significant segregation patterns between OTUs (Louca et al., 2016). In both the present study and in the bromeliads, the actual biological mechanisms driving this segregation remain unknown and may include, for example, apparent competition driven by phages and antibiotic warfare between bacteria (Louca et al., 2017; Louca et al., 2018). That said, additional experimental evidence, such as direct observations of interactions or experiments manipulating community composition, is needed to confirm these statistical inferences

2.3 Discussion:

Our findings as described above, revealed that, across a large-scale transect of subsurface coastal sediments, the functional structure of microbial communities remained stable despite significant variations in taxonomic composition at lower levels. This suggests that functional redundancy maintains consistent functionality even with different taxa fulfilling similar roles. Environmental filtering influences metabolic functions across depths, but additional mechanisms may contribute to high taxonomic variability. Our analysis showed limited evidence that geographic distance or environmental factors drive this variation, while biological interactions, particularly antagonistic ones, seem to play a major role in community assembly. This suggests a separation between functional structure and taxonomic diversity within functional groups, each shaped by distinct processes. Further investigation is needed to better understand these mechanisms, especially the role of biological interactions, through experimental and broader field-based studies. This could help refine models for microbial community assembly by separating functional and taxonomic variation.

3. CONCLUSSION

This chapter includes material co-authored by me and Dr. Stilianos Louca. It has been submitted for publication in a peer-reviewed journal, along with contributions from other co-authors, whose work in earlier chapters has been acknowledged.

In this study, we have presented a systematic examination of the microbial communities in subsurface coastal sediments, along a transect spanning hundreds of kilometers and covering multiple river deltas. Despite these large spatial scales, the functional structure of the communities in terms of the proportions of various gene groups was remarkably constant across all samples, even at the highest resolutions considered (ECs and KEGG level C). In contrast to this stable functional structure, we observed strong variations in taxonomic composition, especially at lower taxonomic levels (ASV, OTU, and genus).

Overall, these results suggest that community-level function in marine subsurface sediments may be decoupled from taxonomic composition within functional groups. In particular, despite the strong environmental filters of metabolic functions across depth (Urakawa et al., 2000; Edlund et al., 2008; Durbin & Teske, 2011; Molari et al., 2012; Vuillemin et al., 2018), additional mechanisms can cause high taxonomic variation within a given layer. Conceptually, this means that one might separate community composition into two perpendicular axes of variation: function on one hand and taxonomic composition within functional groups on the other hand, with each axis being controlled by separate mechanisms (Louca et al., 2016; Louca, 2017). Such a separation, in turn, can guide the development of nested models for microbial community assembly—first modeling function regardless of taxonomic composition (Reed et al., 2014; Louca et al., 2022) and subsequently modeling taxonomic variation separately in each functional group (Ofițeru et al., 2010).

We also statistically investigated various alternative factors that might explain the observed taxonomic variation despite the strongly constrained functional structure, including geographic distance, several environmental variables, and intrinsic biological interactions. We found little evidence that geographic distance and the considered environmental variables drive this variation at the considered spatial scales. Instead, we conclude that biological interactions—primarily antagonistic—likely strongly affect community assembly and appear to cause a statistically significant segregation of the distributions of various taxa across samples.

APPENDIX: SUPPLEMENTAL MATERIAL

Table S 1 Overview of Samples general information. Sampling locations, sampling dates and numbers of raw metagenomic and 16S rRNA gene amplicon reads.

sample	nearest delta	date (Y.M.D)	latitude	longitude	metag. reads	16S reads
ColS0	Columbia	2023.10.14	46.224290	-124.009156	2979797	10499
CsN0	Coos	2023.10.21	43.357270	-124.338938	6455597	18295
CsN2	Coos	2023.10.21	43.372205	-124.326941	3958666	18975
CsN4	Coos	2023.10.21	43.389272	-124.314713	4492886	6910
NN35	Nehalem	2023.11.11	45.689080	-123.940613	5319442	6113
NN5	Nehalem	2023.11.11	45.703327	-123.941195	4210128	11504
SilS5	Siletz	2023.11.12	44.882809	-124.036913	5810252	9234
SiuS0	Siuslaw	2023.10.08	44.016103	-124.138902	6891881	25207
SiuS2	Siuslaw	2023.10.08	43.998559	-124.138189	5555930	10920
SiuS4	Siuslaw	2023.10.08	43.980818	-124.140290	2209106	801

Table S 2 Overview of the major river deltas near the sampling locations (United States Geological Survey, 2007, 2010a, 2010b, 2010c; D'Hondt et al., 2004).

delta	typical discharge rate	origin
Columbia	5500 $m^3 \cdot s^{-1}$	Rocky Mountains of BC, Canada
Coos	41 $m^3 \cdot s^{-1}$	South-West Oregon
Nehalem	75 $m^3 \cdot s^{-1}$	Northern Oregon coast range
Siletz	42 $m^3 \cdot s^{-1}$	Central Oregon coast range
Siuslaw	56 $m^3 \cdot s^{-1}$	Central Oregon coast range

Table S 3 Mantel tests (taxonomic dissimilarities vs geographic distances). Overview of Mantel tests of Spearman rank correlations, at each taxonomic level and for each of the three considered dissimilarity metrics. The table lists Spearman correlation coefficients (ρ) and associated one-sided statistical significances (P). Significances below 0.05 are bolded. Adjusting the significance threshold to account for multiple hypothesis tests using a Bonferroni correction ($\alpha = 0.05/21 = 0.0024$) renders all Mantel tests statistically insignificant.

	Jaccard		**Bray-Curtis**		**Hellinger**	
tax. level	ρ	P	ρ	P	ρ	P
phylum	0.34	**0.01**	-0.13	0.24	0.38	**0.015**
class	-0.043	0.40	-0.027	0.48	0.28	0.057
order	-0.055	0.39	-0.13	0.26	0.21	0.12
family	0.27	0.095	-0.10	0.31	0.27	0.052
genus	0.28	**0.031**	-0.037	0.42	0.24	0.089
OTU	0.23	0.057	0.024	0.39	0.13	0.17
ASV	-0.057	0.37	-0.02	0.50	0.16	0.20

Table S 4 Regression models. Overview of regression model selection, predicting pairwise taxonomic dissimilarities (Jaccard, Bray-Curtis or Hellinger) as a function of pairwise absolute differences in environmental variables and/or geographic distances. The table includes the predictors selected at each taxonomic level and associated P-values in parentheses, as well as the fraction of variance explained by the models based on leave-one-out cross validation (Rcv 2). Note that in most cases no predictors were selected, i.e., none had a statistically significant coefficient. In fact, if the significance threshold were to be adjusted for multiple hypothesis tests ($\alpha = 0.05/(3 \cdot 7 \cdot 21) = 0.00011$), then no predictor would have been selected in any case.

	Jaccard		**Bray-Curtis**		**Hellinger**	
tax. level	**predictors**	R^2_{cv}	**predictors**	R^2_{cv}	**predictors**	R^2_{cv}
phylum	-	-	-	-	Boron (P=0.042)	0.27
class	-	-	-	-	Boron (P=0.036)	0.26
order	-	-	-	-	-	-
family	-	-	-	-	Boron (P=0.033)	0.22
genus	-	-	-	-	Boron (P=0.023)	0.24
OTU	-	-	-	-	-	-
ASV	-	-	-	-	-	-

Table S 5 Cooccurrence analysis (CC scores). Overview of checkerboard cooccurrence (CC) analyses across samples, at various taxonomic levels. The table shows the number of taxa analyzed, the CC score of taxon presences/absences, the mean CC score under the null model (i.e., expectation under independent taxon distributions), the standardized effect size (SES) and the two-sided statistical significance (probability that the null model would generate a more extreme SES than observed). Note that a lower CC score implies a lower overlap of taxon distributions. Taxonomic levels where the CC score was significantly different from the null model's expectation (P<0.05) are bolded. Taxonomic levels for which P falls below the Bonferroni-corrected significance threshold (P<0.0071), i.e., accounting for 7 independent hypothesis tests, are underlined.

tax. level	Ntaxa	CC score	mean null	SES	significance (P)
phylum	34	0.623	0.635	-0.53	0.563
class	75	0.624	0.659	-2.60	0.011
order	130	0.443	0.489	-6.14	<0.001
family	155	0.392	0.441	-7.41	<0.001
genus	193	0.397	0.450	-10.5	<0.001
OTU	1507	0.275	0.311	-32.1	<0.001
ASV	4315	0.234	0.253	-36.1	<0.001

Table S 6 Co-abundance analysis (MA scores). Overview of the Morisita coabundance analyses across samples, at various taxonomic levels. The table shows the number of taxa analyzed, the MA score of relative taxon abundances, the mean MA score under the null model (i.e., expectation under independent taxon distributions), the standardized effect size (SES) and the two-sided statistical significance (probability that the null model would generate a more extreme SES than observed). Note that a lower MA score implies a lower overlap of taxon distributions. At all taxonomic levels, MA scores were significantly lower than the null model's expectation (P<0.05), even after a Bonferroni correction accounting for multiple hypothesis tests (P<0.0071).

tax. level	Ntaxa	MA score	mean null	SES	significance (P)
phylum	34	0.922	0.998	-81.6	<0.001
class	75	0.896	0.996	-79.8	<0.001
order	130	0.872	0.994	-69.6	<0.001
family	155	0.884	0.994	-62.4	<0.001
genus	193	0.830	0.984	-47.7	<0.001
OTU	1507	0.459	0.844	-60.0	<0.001
ASV	4315	0.225	0.472	-44.8	<0.001

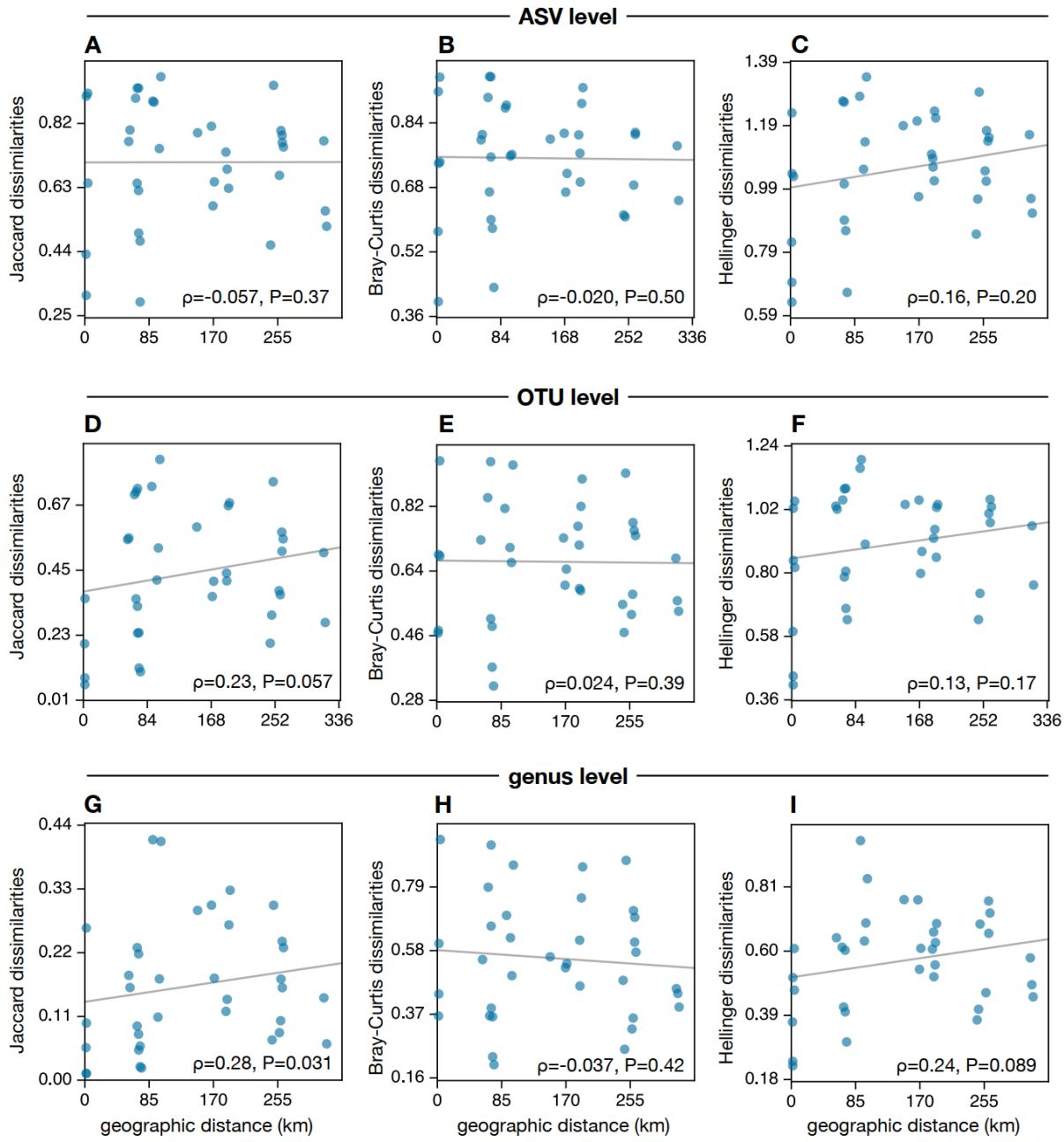

Figure S. 1 Mantel Tests. (A) Pairwise Jaccard dissimilarities between samples at the ASV level (vertical axis) compared to pairwise geographic distances (horizontal axis). Each point represents a distinct pair of samples. A linear regression line (fitted via least squares) is shown for reference. The Spearman rank correlation (ρ) and its statistical significance (P), assessed via a Mantel test (random permutations of the rows and columns of the dissimilarity matrix), are shown inside the figure. (B, C) Similar to A, but considering the Bray-Curtis and Hellinger dissimilarity metrics. (D–F) Similar to A–C but evaluating dissimilarities at the level of OTUs. (G–I) Similar to A–C but evaluating dissimilarities at the level of genera. For additional taxonomic levels, see Supplemental Table S3.

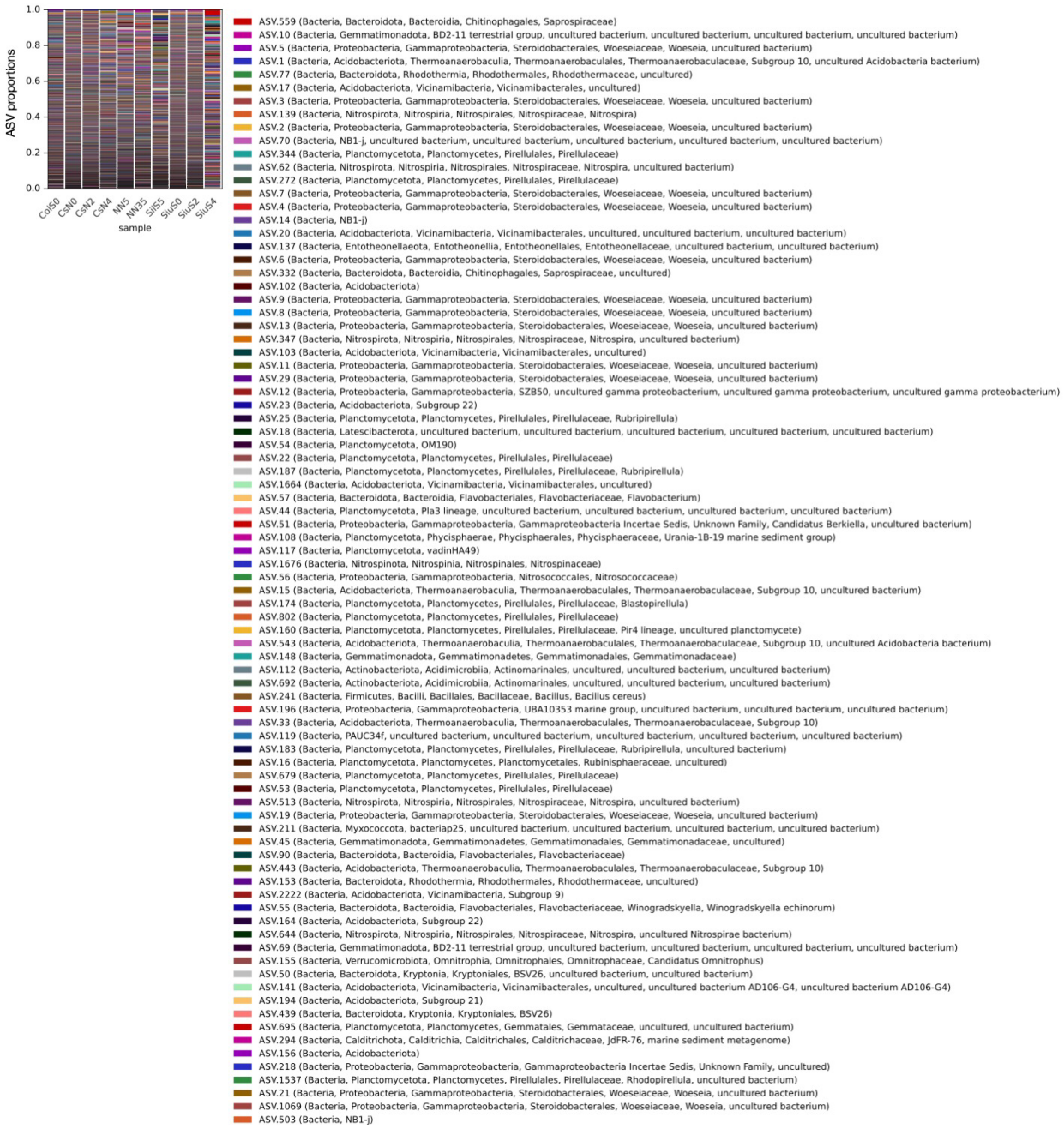

Figure S. 2 ASV proportions. Proportions (relative abundances) of 16S rRNA gene amplicon sequence variants (ASVs) detected in all samples, based on the number of reads mapped to each ASV. ASVs are sorted from top to bottom in decreasing average proportion. Only the most abundant ASVs are listed in the legend for readability. Estimated taxonomic identities are shown in parentheses. While archaeal ASVs were also detected, they were relatively rare and are thus not listed in the legend.

BIBLIOGRAPHY:

Anderson, R. E. (2021). Tracking microbial evolution in the subseafloor biosphere. *mSystems, 6*(4), e00731-21. https://doi.org/10.1128/msystems.00731-21

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., & Ogata, H. (2019). KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics, 36*(2), 2251–2252. https://doi.org/10.1093/bioinformatics/btz859

Burke, C., Steinberg, P., Rusch, D., Kjelleberg, S., & Thomas, T. (2011). Bacterial community assembly based on functional genes rather than species. *Proceedings of the National Academy of Sciences, 108*(34), 14288–14293. https://doi.org/10.1073/pnas.1101591108

Caldwell, D. R. (1973). Thermal and Fickian diffusion of sodium chloride in a solution of oceanic concentration. *Deep Sea Research and Oceanographic Abstracts, 20*(12), 1029–1039. https://doi.org/10.1016/0011-7471(73)90073-9

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods, 13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Canfield, D. E., & Thamdrup, B. (2009). Towards a consistent classification scheme for geochemical environments, or, why we wish the term 'suboxic' would go away. *Geobiology, 7*(4), 385–392. https://doi.org/10.1111/j.1472-4669.2009.00214.x

Chao, A., Chazdon, R. L., Colwell, R. K., & Shen, T. J. (2005). A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology Letters, 8*(2), 148–159. https://doi.org/10.1111/j.1461-0248.2004.00707.x

Czárán, T. L., Hoekstra, R. F., & Pagie, L. (2002). Chemical warfare between microbes promotes biodiversity. *Proceedings of the National Academy of Sciences, 99*(12), 786–790. https://doi.org/10.1073/pnas.012399899

D'Hondt, S., Jørgensen, B. B., Miller, D. J., Batzke, A., Blake, R., Cragg, B. A., Cypionka, H., Dickens, G. R., Ferdelman, T., Hinrichs, K.-U., Holm, N. G., Mitterer, R., Spivack, A., Wang, G., Bekins, B., Engelen, B., Ford, K., Gettemy, G., Rutherford, S. D., ... & Southam, G. (2004). Distributions of microbial activities in deep subseafloor sediments. *Science, 306*(5705), 2216-2221. https://doi.org/10.1126/science.1101155

Durbin, A. M., & Teske, A. (2011). Microbial diversity and stratification of South Pacific abyssal marine sediments. *Environmental Microbiology, 13*(12), 3219–3234. https://doi.org/10.1111/j.1462-2920.2011.02544.x

Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bty113

Edlund, A., Hårdeman, F., Jansson, J. K., & Sjöling, S. (2008). Active bacterial community structure along vertical redox gradients in Baltic Sea sediment. *Environmental Microbiology, 10*(8), 2051–2063. https://doi.org/10.1111/j.1462-2920.2008.01624.x

Falkowski, P. G., Fenchel, T., & Delong, E. F. (2008). The microbial engines that drive Earth's biogeochemical cycles. *Science, 320*(5879), 1034–1039. https://doi.org/10.1126/science.1153213

Fernandez, A. S., Hashsham, S. A., Dollhopf, S. L., Raskin, L., Glagoleva, O., Dazzo, F. B., & Criddle, C. S. (2000). Flexible community structure correlates with stable community function in methanogenic bioreactor communities perturbed by glucose. *Applied and Environmental Microbiology, 66*(9), 4058–4067. https://doi.org/10.1128/aem.66.9.4058-4067.2000

Fernández, A., Huang, S. Y., Seston, S., Xing, J., Hickey, R., Criddle, C., & Tiedje, J. (1999). How stable is stable? Function versus community composition. *Applied and Environmental Microbiology, 65*(8), 3697–3704. https://doi.org/10.1128/aem.65.8.3697-3704.1999

Glöckner, F. O., Yilmaz, P., Quast, C., Yarza, P., Peplies, J., & Ludwig, W. (2017). 25 years of serving the community with ribosomal RNA gene reference databases and tools. *Journal of Biotechnology, 261*, 169–176. https://doi.org/10.1016/j.jbiotec.2017.06.1198

Hoehler, T. M., & Jørgensen, B. B. (2013). Microbial life under extreme energy limitation. *Nature Reviews Microbiology, 11*(2), 83–94. https://doi.org/10.1038/nrmicro2939

Hoshino, T., Inagaki, F., & Ijiri, A. (2020). Global diversity of microbial communities in marine sediment. *Proceedings of the National Academy of Sciences, 117*(46), 27587–27597. https://doi.org/10.1073/pnas.1919139117

Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics, 11*, 119. https://doi.org/10.1186/1471-2105-11-119

Iversen, N., & Jørgensen, B. B. (1993). Diffusion coefficients of sulfate and methane in marine sediments: Influence of porosity. *Geochimica et Cosmochimica Acta, 57*(3), 571–578. https://doi.org/10.1016/0016-7037(93)90368-7

Jiang, H., Lei, R., Ding, S. W., & Zhu, S. (2014). Skewer: A fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics, 15*, 1–12. https://doi.org/10.1186/1471-2105-15-182

Jørgensen, B. B. (2011). Deep subseafloor microbial cells on physiological standby. *Proceedings of the National Academy of Sciences, 108*(46), 18193–18194. https://doi.org/10.1073/pnas.1115421108

Kabay, N., Bryjak, M., & Hilal, N. (2015). *Boron separation processes*. Elsevier.

Kallmeyer, J., Pockalny, R., Adhikari, R. R., Smith, D. C., & D'Hondt, S. (2012). Global distribution of microbial abundance and biomass in subseafloor sediment. *Proceedings of the National Academy of Sciences, 109*(40), 16213–16216. https://doi.org/10.1073/pnas.1203849109

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., & Tanabe, M. (2021). KEGG: integrating viruses and cellular organisms. *Nucleic Acids Research, 49*(D1), D545–D551. https://doi.org/10.1093/nar/gkaa970

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., & Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Research, 32*(D1), D277–D280. https://doi.org/10.1093/nar/gkh063

Kim, M., Oh, H. S., Park, S. C., & Chun, J. (2014). Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology, 64*(2), 346–351. https://doi.org/10.1099/ijs.0.059774-0

Legendre, P., & Gallagher, E. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia, 129*(2), 271–280. https://doi.org/10.1007/s004420100716

Legendre, P., & Legendre, L. (1998). *Numerical ecology* (2nd ed.). Elsevier.

Legendre, P., Lapointe, F. J., & Casgrain, P. (1994). Modeling brain evolution from behavior: A permutational regression approach. *Evolution, 48*(5), 1487–1499. https://doi.org/10.1111/j.1558-5646.1994.tb02191.x

Li, D., Liu, C. M., Luo, R., Sadakane, K., & Lam, T. W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics, 31*(10), 1674–1676. https://doi.org/10.1093/bioinformatics/btv033

Louca, S. (2017). Probing the metabolism of microorganisms. *Science, 358*(6369), 1264–1265. https://doi.org/10.1126/science.aar2000

Louca, S. (2021). The rates of global bacterial and archaeal dispersal. *ISME Journal, 16*(1), 159–167. https://doi.org/10.1038/s41396-021-01069-8

Louca, S., Doebeli, M., & Parfrey, L. W. (2017). Taxonomic variability and functional stability in microbial communities infected by phages. *Environmental Microbiology, 19*(9), 3863–3878. https://doi.org/10.1111/1462-2920.13743

Louca, S., Mazel, F., Doebeli, M., & Parfrey, L. W. (2020). Effects of forced taxonomic transitions on metabolic structure and function in microbial microcosms. *Environmental Microbiology Reports, 12*(4), 514–524. https://doi.org/10.1111/1758-2229.12866

Louca, S., Parfrey, L. W., & Doebeli, M. (2016). Decoupling function and taxonomy in the global ocean microbiome. *Science, 353*(6305), 1272–1277. https://doi.org/10.1126/science.aaf4507

Louca, S., Polz, M. F., Mazel, F., Albright, M. B. N., Huber, J. A., & O'Connor, M. I. (2018). Function and functional redundancy in microbial systems. *Nature Ecology & Evolution, 2*(6), 936–943. https://doi.org/10.1038/s41559-018-0519-1

Louca, S., Taylor, G. T., Astor, Y. M., Buck, K., & Muller-Karger, F. E. (2022). Transport-limited reactions in microbial systems. *Environmental Microbiology, 25*(1), 268–282. https://doi.org/10.1111/1462-2920.16275

Lutz, G., Hubbell, D., Stevens, H., (U.S.), G. S., & Commission, U. A. E. (1975). *Discharge and flow distribution, Columbia River estuary*. U.S. Government Printing Office.

Magoc, T., & Salzberg, S. L. (2011). FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics, 27*(21), 2957–2963. https://doi.org/10.1093/bioinformatics/btr507

Molari, M., Giovannelli, D., d'Errico, G., & Manini, E. (2012). Factors influencing prokaryotic community structure composition in sub-surface coastal sediments. *Estuarine, Coastal and Shelf Science, 97*, 141–148. https://doi.org/10.1016/j.ecss.2011.11.036

Nelson, M. B., Martiny, A. C., & Martiny, J. B. H. (2016). Global biogeography of microbial nitrogen-cycling traits in soil. *Proceedings of the National Academy of Sciences, 113*(29), 8033–8040. https://doi.org/10.1073/pnas.1601070113

Nealson, K. H. (1997). Sediment bacteria: Who's there, what are they doing, and what's new? *Annual Review of Earth and Planetary Sciences, 25*, 403–434. https://doi.org/10.1146/annurev.earth.25.1.403

Ofițeru, I. D., Lunn, M., Curtis, T. P., Wells, G. F., Criddle, C. S., Francis, C. A., & Sloan, W. T. (2010). Combined niche and neutral effects in a microbial wastewater treatment community. *Proceedings of the National Academy of Sciences, 107*(35), 15345–15350. https://doi.org/10.1073/pnas.1000604107

Palmer, T., & Avery, W. (2014). *Field guide to Oregon rivers*. Oregon State University Press.

Parada, A. E., Needham, D. M., & Fuhrman, J. A. (2016). Every base matters: Assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental Microbiology, 18*(5), 1403–1414. https://doi.org/10.1111/1462-2920.13023

Parkes, R., Cragg, B., & Wellsbury, P. (2000). Recent studies on bacterial populations and processes in subseafloor sediments: A review. *Hydrogeology Journal, 8*(1), 11–28. https://doi.org/10.1007/PL00010971

Reagan, J., Gregg, W., Levine, P., & Mishonov, A. (2024). *World Ocean Atlas 2023: Product documentation*. NOAA Ocean Climate Laboratory.

Reed, D. C., Algar, C. K., Huber, J. A., & Dick, G. J. (2014). Gene-centric approach to integrating environmental genomics and biogeochemical models. *Proceedings of the National Academy of Sciences, 111*(5), 1879–1884. https://doi.org/10.1073/pnas.1313713111

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open-source tool for metagenomics. *PeerJ, 4*, e2584. https://doi.org/10.7717/peerj.2584

Schippers, A., Kock, D., Höft, C., Köweker, G., & Siegert, M. (2012). Quantification of microbial communities in subsurface marine sediments of the Black Sea and off Namibia. *Frontiers in Microbiology, 3*, 16.

Soufi, H. H., Tran, D., & Louca, S. (2024). Microbiology of Big Soda Lake, a multi-extreme meromictic volcanic crater lake in the Nevada desert. *Environmental Microbiology, 26*(4), e16578. https://doi.org/10.1111/1462-2920.16578

Teske, A. P. (2006). Microbial Communities of Deep Marine Subsurface Sediments: Molecular and Cultivation Surveys. *Geomicrobiology Journal, 23*(6), 357–368. https://doi.org/10.1080/01490450600875613

Thorn, P. M., & Ventullo, R. M. (1988). Measurement of bacterial growth rates in subsurface sediments using the incorporation of tritiated thymidine into DNA. *Microbial Ecology, 16*(1), 3–16. https://doi.org/10.1007/bf02097401

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2009). A core gut microbiome in obese and lean twins. *Nature, 457*(7228), 480–484. https://doi.org/10.1038/nature07540

United States Geological Survey. (2007). *Water-data report 2007: 14301000 Nehalem River near Foss, OR*. United States Geological Survey.

United States Geological Survey. (2010). *Water-Data Report 2010: 14305500 Siletz River at Siletz, OR*. United States Geological Survey.

United States Geological Survey. (2010). *Water-Data Report 2010: 14307620 Siuslaw River near Mapleton, OR*. United States Geological Survey.

Urakawa, H., Yoshida, T., Nishimura, M., & Ohwada, K. (2000). Characterization of depth-related population variation in microbial communities of a coastal marine sediment using 16S rDNA-based approaches and quinone profiling. *Environmental Microbiology, 2*(5), 542–554. https://doi.org/10.1046/j.1462-2920.2000.00137.x

Vuillemin, A., Wörmer, L., Alawi, M., Henny, C., Wagner, D., & Orsi, W. D. (2018). Microbial community composition along a 50,000-year lacustrine sediment sequence. *FEMS Microbiology Ecology, 94*(4), fiy029. https://doi.org/10.1093/femsec/fiy029

Walters, W., Hyde, E. R., Berg-Lyons, D., Ackermann, G., Humphrey, G., Parada, A., Gilbert, J. A., Jansson, J. K., Caporaso, J. G., & Knight, R. (2015). Improved bacterial 16S rRNA gene (V4 and V4-5) and fungal internal transcribed spacer marker gene primers for microbial community surveys. *mSystems, 1*(1), e00009-15. https://doi.org/10.1128/mSystems.00009-15

Williams, L. B., Hervig, R. L., Wieser, M. E., & Hutcheon, I. (2001). The influence of organic matter on the boron isotope geochemistry of the Gulf Coast sedimentary basin, USA. *Chemical Geology, 174*(1-3), 445–461. https://doi.org/10.1016/S0009-2541(00)00289-8