# THEORETICAL AND EMPIRICAL METHODS CONNECTING MICROBIAL METABOLISM DIVERSITY WITH MICROBIAL PHYLOGENETICS, COMMUNITY DIVERSITY, AND ORGANIC GEOCHEMISTRY

**A Dissertation Presented for the**
**Doctor of Philosophy**
**Degree**
**The University of Tennessee, Knoxville**

**Taylor Michael Royalty**
**August 2020**

# DEDICATION

I dedicate this to my daughter and wife.

## ACKNOWLEDGEMENTS

For Chapter One, I thank Chad Burdyshaw of JICS for help obtaining the genomes used in this project and C-DEBI for funding my graduate fellowship. For Chapter Two, I thank C-DEBI for funding my graduate fellowship. For Chapter Three, I thank Dr. Karen Lloyd for providing feedback on the original draft of this chapter as well as C-DEBI for funding my graduate fellowship. For Chapter Four, I thank Dr. Anthony Faiia from the UTK Stable Isotope Lab for providing expertise on ion chromatography and isotope measurements, Roberto Diaz for making methane measurements, Dr. Karen Lloyd for providing insights about geochemistry at the White Oak River estuary, and C-DEBI for funding my graduate fellowship. I thank my Ph.D. committee, comprised of Dr. Annette Engel, Dr. Robert Hettich, and Dr. Anna Szynkiewicz, for their academic guidance. Lastly, I thank Drew for guiding me through all aspects of the Ph.D.

# ABSTRACT

Biogeochemistry is controlled by microorganisms obtaining nutrients and energy. Thus, microbial metabolisms directly link microbial ecology and geochemistry. The extent that microbial ecology and geochemistry microbial ecology and geochemistry affects the other requires constraint on the spatiotemporal distribution and abundance of microbial metabolisms with respect to geochemistry, or the microbial niches. Elucidating microbial metabolisms was challenging prior to the advent of ‘omics sequencing technologies, as most microbial lineages lack cultured representatives. Although revolutionizing microbial ecology, challenges still exist in fully leveraging information derived from ‘omics technologies. This dissertation attempts to address a small subset of these challenges that include quantifying the generalizability of microbial metabolism with respect to phylogeny, relating metagenomic sequencing effort to *in situ* genome discovery rates, quantifying and generalizing the relative contribution to a net ecosystem function by community members, and relating geochemistry gradients to microbial metabolism gradients. As a part of this work, theoretical and quantitative measures are proposed for evaluating microbial metabolism diversity with respect to phylogenetics (permutational multivariate ANOVA and variance component modeling), community diversity (generalized coupon collector equation, parametric diversity), and *in situ* geochemistry at the field site, White Oak River estuary, North Carolina (USA). Numerical simulations (community rarefaction, community extinction events, and reaction-transport modeling) and public data repositories (Reference Sequence Database, GenBank, Integrated Microbial Genome and Microbiomes, and Sequence Read Archive) are used for the testing efficacy of the proposed theoretical and quantitative methods. The results indicate that numerical simulations and public data repositories

can be used for developing and testing ecological theory and concepts. The theoretical and quantitative methods proposed here can now be used in exploring microbial niche distributions in nature.

# TABLE OF CONTENTS


Introduction ..... 1
Chapter One Quantitatively Partitioning Microbial Genomic Traits among Taxonomic Ranks Across the Microbial Tree of Life ..... 7
Abstract ..... 8
Importance ..... 9
Introduction ..... 9
Results ..... 11
Discussion ..... 16
Materials and Methods ..... 20
Genome Database Curation ..... 20
COG Functional Category Identification, Enumeration, and Normalization ..... 21
Principal Component Analysis (PCA) ..... 22
Quantifying COG-FC Variance Explained by Taxonomic Rank ..... 22
Appendix One ..... 25
Tables ..... 26
Figures ..... 29
Chapter Two Theoretical and Simulation-based Investigation of the Relationship between Sequencing Effort, Microbial Community Richness, and Diversity in Binning Metagenome-assembled Genomes ..... 36
Abstract ..... 37
Importance ..... 38
Introduction ..... 38
Results ..... 40
Theoretical and Numerical Sequencing Effort Simulations ..... 40
Rarefying MAG Binning as a function of sequencing effort ..... 43
Constraining MAG Rarefaction Analyses to Community Structure ..... 44
Discussion ..... 45
Materials and Methods ..... 50
Defining the Microbial Metagenome and Sequencing Probability ..... 50
Modeling Expected Sequences ..... 52
Sequence Data Sources ..... 55
MAG Assembly Pipeline ..... 56
Subsampling Sequence Read Datasets ..... 57
Modeling MAG Response to Sequencing effort ..... 58
Relating MAG Response to the Theoretical Sequencing Model ..... 58
Appendix Two ..... 60
Tables ..... 61
Chapter Three A Quantitative Measure of Functional Redundancy in Microbial Ecosystems ..... 71
Abstract ..... 72
Importance ..... 73
Introduction ..... 73
Results and Discussion ..... 76

Representing microbial communities as taxa abundance, trait abundance, and trait contribution distributions ..... 76
Defining and Quantifying Functional Redundancy ..... 77
Influence on Functional Redundancy from Taxa Lacking the Trait ..... 79
Functional Redundancy and Trait Resilience to Taxa Extinction ..... 82
Case Study: Redundancy of Nitrogen Cycle Marker Genes in Marine Microbiomes ..... 84
Methods and Theory ..... 86
Definition of Functional Redundancy ..... 86
Functional Redundancy's Dependency on Diversity Continuum ..... 89
Functional Redundancy Against Trait Vulnerability Simulations ..... 90
A Case Study of Functional Redundancy in the Global Ocean Nitrogen Cycle ..... 91
Appendix Three ..... 93
Tables ..... 94
Figures ..... 95
Chapter Four Sulfur and Carbon Budgets in the White Oak River Estuary Based on Reaction-Transport Modeling ..... 101
Abstract ..... 102
Importance ..... 103
Introduction ..... 104
Materials and Methods ..... 105
Site Description ..... 105
Core Sampling ..... 107
Methane Analysis ..... 107
Porewater Chemistry ..... 108
Ion Chromatography ..... 109
DOM Absorption Spectra ..... 110
Total Organic Carbon, Total Organic Nitrogen, and Bulk $\delta^{13}C$ ..... 111
Reaction-Transport Model ..... 113
Total Organic Carbon, Methane, and Sulfur Budgets ..... 115
Castle Hayne Aquifer Sulfate Concentrations ..... 116
Results ..... 116
Porewater Ion Depth Profiles ..... 116
Castle Hayne Aquifer Ion Concentrations ..... 117
Organic Carbon Depth Profiles ..... 118
Groundwater Upwelling and Sulfate/Methane Production Rates ..... 119
Carbon and Sulfur Budgets in Station H White Oak River Estuary Sediments ..... 120
Discussion ..... 121
Castle Hayne Aquifer and Ion Dilution in WOR Sediments ..... 122
Metabolic Alterations to Organic Carbon in WOR Sediments ..... 123
Evaluating the Carbon and Sulfur Budgets at WOR, Station H ..... 125
Conclusions and Future Directions ..... 127
Appendix Four ..... 129
Tables ..... 130
Figures ..... 132

Conclusions ..... 141
List of References ..... 144
Vita ..... 167

# LIST OF TABLES

Table 1.1. A summary of the custom-curated genome database used in this work. ..... 26
Table 1.2. Fit statistics for GAM regressions modeling COG-FC abundance as a function of genome size. ..... 27
Table 1.3. Phylum highly enriched (>85$^{th}$ percentile) or depleted (<15$^{th}$ percentile) in COG-FCs and depletion. All reported categories are statistically significant. .....28
Table 2.1. Estimates of fit coefficients for the Gompertz equation (equation 2.12) for the sum of high-quality and medium-quality MAGs as a function of sequencing depth in published data sets from ocean surface water, estuarine sediment, maize soil, and the human gut* .....61
Table 2.2. Summary of sequence datasets analyzed with the MAG pipeline. .....62
Table 3.1. A summary of marker gene sequences used in the BLASTP analysis of *TARA Oceans* MAGs. All marker genes were downloaded from Swiss-Prot. .....94
Table 4.1. Pairwise slopes for porewater ion concentrations. All slopes are significant ($p<0.001$). .....130
Table 4.2. A summary of carbon and sulfur budgets in sediments at WOR estuary, Station H. .....131

# LIST OF FIGURES

Figure 1.1. Violin plots showing the distribution for the ratio of total COG-FC annotations in a genome to the total number of open reading frames for each phylum. ..... 29
Figure 1.2. PCA plots of COG-FC abundances (A) and relative abundances (B and C). Individual data points are colored by genome size in panels A and B. The data presented in panel A were not normalized by genome size, while the data in panels B and C were normalized by genome size. Black contours on panels B and C correspond to density plots for all genomes shown in panel B. Colored contours in panel C correspond to the respective lineage labels. For panel A, PC1 explained 71% and PC2 explained 7.0% of the variance. For panels B and C, PC1 explained 21% and PC2 explained 16% of the variance. Panel C corresponds to only the top 10 most abundant phyla analyzed as described for Table 1.1..... 30
Figure 1.3. GAM regressions modeling COG-FC normalized abundance (standardized) as a function of genome size. Solid red lines correspond to mean fit. Upper and lower red dashed lines correspond to 95th percentile confidence intervals. ..... 31
Figure 1.4. The average variance in COG-FC enrichment explained by different taxonomic ranks (bars) and the cumulative variance explained by taxonomic ranks (lines). All variance explained by taxonomic ranks was significant ($p$<0.001). The $F$-value for domain, phylum, class, order, family, and genus, was 726.0, 128.8, 38.76, 34.4, 11.2, and 5.1, respectively. ..... 32
Figure 1.5. Violin plots showing the distribution in $\log_{10}$-distance for lineages from their respective phyla mean COG-FC centroid (A) and average $\log_{10}$-distance to phyla mean COG-FC centroid as a function of number of lineages in a given phylum (B). Coefficients in panel B correspond to fit parameters from eq 1. The * symbols denote significance as defined in the text. We note three outliers: the *Crenarhaeota* are characterized by unusually high diversity of COG-FCs distributions, and the *Synergistota* and *Fibrobacterota* are characterized by an unusually low diversity of COG-FC distributions. ..... 33
Figure 1.6. A heat map showing the average COG-FC enrichment for all archaeal (A) and bacterial (B) genera. Categories were arranged from left to right along the x axis in order of decreasing total variance in enrichment across all lineages. Clades were organized along the y axis using phylogenetic relatedness based on the reported concatenated protein sequence alignments in (10). ..... 34
Figure 1.7. Results from a variance component model. Lineage was used as a nested random effect (intercept) for all COG-FCs. The proportion of variance explained is partitioned by phylum (A), class (B), order (C), family (D), and genus (E). Box plots correspond to the variability in variance explained from the bootstrap analysis, red dashed lines correspond to 95% confidence intervals calculated from the bootstrap analysis, and red circles correspond to the variance explained by analysis of all data in Table 1.1. Note that the titles for COG-FCs are shortened; the full category names are shown in Figure 1.6 and Table 1.2. 35
Figure 2.1. A flow diagram illustrating the workflow for sequencing experiments. ..... 63
Figure 2.2. Sequencing effort, with units of log of sequence reads, required to fully sequence four different community structures, one with relatively high community evenness (A), relatively moderate community evenness (B), relatively low community evenness (C), and

one with a lognormal community structure (D), were predicted using linear regressions (E) and the log of metagenome size (total possible *k*-mers) as a predictor. ..... 64
Figure 2.3. Sequencing effort, with units of log sequence reads, necessary to reach variable target sequencing depths (colors) for four different community structures, one with relatively high community evenness (A), relatively moderate community evenness (B), relatively low community evenness (C), and one with a lognormal community structure (D). Red translucent lines correspond with linear regression curves for the respective community in Figure 2.3E..... 65
Figure 2.4. Numerical sequencing simulations applied to 6 hypothetical communities with different lognormal distributions that were defined by the parameter, *a*, from equation 2.7 (A). The sequencing effort, with units of log of sequence reads, necessary to sequence a target fraction of exhaustion (dashed contours) as a function of the Pielou evenness index, *J*, for a given lognormal community structure (B). ..... 66
Figure 2.5. Numerical sequencing simulations show the number of bases (color bar) required to sequence a target fraction of a genome as a function of that genome's relative abundance in the community metagenome. Genomes evaluated were $0.5\times10^6$ (A), $2\times10^6$ (B), $5\times10^6$ (C), $10\times10^6$ (D), and $20\times10^6$ (E) base pairs long..... 67
Figure 2.6. The influence that the parameters $A$, $\mu$, and $\lambda$ had on the Gompertz equation (A). The property of the Gompertz equation that each parameter influences is colored red. The sum of medium quality and high quality (quality MAGs) as a function of sequencing effort for the gut (B), soil (C), surface ocean (D), and sediment (E) communities. Solid lines in (B-E) correspond to nonlinear least squares fits of the Gompertz equation to the respective environmental dataset. ..... 68
Figure 2.7. Quality MAGs (medium and high quality) as a function of rarest genome sequenced to 50% exhaustion for human gut, maize soil, estuarian sediment, and surface ocean sequence datasets. Sequencing effort was converted to genome relative abundance sequenced to target fraction using the GAM presented in Figure 2.5. The translucent shaded areas correspond to uncertainty from the target genome size (1 Mbp, or the lower-bound, to 20 Mbp, or the upper-bound) while the solid lines correspond to genome sizes 5 Mbp. ..... 69
Figure 2.8. A cartoon illustrating an example microbial community ($G$), the metagenomes for genomes ($g_{MG,i}$) (as defined in equation 2.2) within $G$, and the overall metagenome for the given microbial community ($G_{MG}$). In this example, there are 6 genomes ($s$=6) and a total of 13 individual microbes. (A) Black circles represent individual microbes whose genomes are averaged together, $g$. The average genome, $g$, are indicated by different color inner-circles. (B) Individual average genomes can be sequenced at $K$ unique positions depending on the characteristic read length, $k$, of a sequencer. (C) All unique positions that can be sequenced for a given genome, $g$, defines the metagenome, $g_{MG}$, for the $i^{th}$ genome, $g_i$. (D) Replacing all individual genomes in (A) with metagenomes, $g_{MG}$, gives the metagenome of the microbial community, $G_{MG}$..... 70
Figure 3.1. The interplay between environmental parameters and microbial community composition requires an understanding of both community taxa abundances as well as trait levels for respective taxa (A). Community taxa abundance and relative trait level are

scalable and provide a measure for the contribution that different taxa have to the total community level trait (B). ..... 95
Figure 3.2. The sensitivity in functional redundancy in response changing $\alpha$ in equation 3.5 (A). Contours correspond to increasing community richness ($n$) while maintaining the structure of the trait-containing community. Comparisons between contours from (A) when $\alpha$=0.5 and $\alpha$=2 (B) as well as when $\alpha$=0.5 and $\alpha$=1 (C). ..... 96
Figure 3.3. The dynamic range in functional redundancy for randomly generated artificial communities as a function of $\alpha$ selected for equation 3.5. Contours correspond to the fraction of the community containing a trait. Note that at $\alpha$ = 1 and the fraction of community was 20%, the dynamic range was smoothed due to precision error from the log-transformation in equation 3.5 when calculating function redundancy. All other data reflect results from artificial communities. ..... 97
Figure 3.4. A comparison of functional redundancy versus the fraction of community richness that goes randomly extinct prior to a 95% of trait level being lost on the community level. Functional redundancy was calculated with equations 3.1-3.7 (α=0.5) (A) as well as using the method in (91) (B)..... 98
Figure 3.5. Same as Figure 3.4A but functional redundancy was calculated with equations 3.1-3.7 such that α=0.1 (A), α=1 (B), and α=2 (C). ..... 99
Figure 3.6. Boxplots showing functional redundancy calculated for eight nitrogen-transforming pathways in *TARA Oceans* draft MAGs. Different colors correspond to sample depth. Whiskers correspond to 150% of the inter-quantile region. ..... 100
Figure 4.1. A map (Google Earth) showing the White Oak River Estuary and Station H..132
Figure 4.2. A flow diagram showing sample processing and analyzes performed on sediments. ..... 133
Figure 4.3. Porewater concentration profiles (mM) for $Br^-$ (A), $Br^-$ (A), $Ca^{2+}$ (B), $Cl^-$ (C), $K^+$ (D), $Mg^{2+}$ (E), $Na^+$ (F), $NH_4^+$ (G), $PO_4^{3-}$ (H), and $SO_4^{2-}$ (I). ..... 134
Figure 4.4. Organic carbon concentration profiles for $CH_4$ (mmol $cm^{-3}$) (A), $\delta^{13}C_{organic}$ (‰ with respect to VPDB) (B), TOC (mmol $cm^{-3}$) (C), TON (mmol $cm^{-3}$) (D), carbon to nitrogen ratio (E), and DOM spectral slope (F). ..... 135
Figure 4.5. Generalized additive model (GAM) fits for $SO_4^{2-}$ depth profiles (mM) (A) and $CH_4$ (mM) (B) used for reactive-transport modeling..... 136
Figure 4.6. Model results for $SO_4^{2-}$ (A) and $CH_4$ (B) from the reaction-transport model. Contours correspond to different assumed rates (cm $yr^{-1}$) of groundwater upwelling. ..... 137
Figure 4.7. A linear regression fit of TOC loss as a function of time assuming an average sedimentation rate of 0.31 cm $yr^{-1}$ (17, 116, 117)..... 138
Figure 4.8. Column integrated $CH_4$ production (nmol $cm^{-2}$ $d^{-1}$) predicted from the reaction-transport model (Figure 4.4B) as a function of groundwater upwelling (cm $yr^{-1}$). 139
Figure 4.9. Column integrated $SO_4^{2-}$ production (nmol $cm^{-2}$ $d^{-1}$) predicted from the reaction-transport model (Figure 4.4A) as a function of groundwater upwelling (cm $yr^{-1}$). The dashed red line corresponds to the groundwater upwelling velocity where WOR sediments at Station H become a source for $SO_4^{2-}$. ..... 140

# INTRODUCTION

An integral link exists between geochemistry and microbial ecology. Through networks of traits, or metabolisms, microbial life mediates chemical reactions to harvest energy and build biomass from nutrients. It is these reactions that drive biogeochemical fluxes (1). Heterogenous element distributions and speciation necessitates plasticity from microbial life in accessing essential nutrients. Only microbes capable of metabolizing the available nutrients will subsist in an environment. Thus, a prerequisite for linking predictive geochemistry with ecological modeling is a fundamental characterization of the metabolisms driving biogeochemistry (2).

An increase in observed microbial metabolism diversity is a consequence from ongoing research in environmental microbiology (1). Nonetheless, quantitative constraints are lacking for the number of metabolisms that exist and the spatiotemporal distribution of metabolisms. New metabolisms are regularly theorized based on thermodynamics, only to be later discovered in nature. A prominent example is anaerobic oxidation of ammonium (anammox). Hypothesized in the 1970s using thermodynamics, an anammox metabolism was not observed in nature until 2003 (3, 4). A more recent example includes sulfur comproportionation, or the reverse of the sulfur disproportionation metabolism, which is thermodynamically favorable at low temperatures and low pH values (5). The theoretical diversity for microbial metabolisms is seemingly boundless, with the primary constraints being the number of chemical compounds and the physical limitations of life (nutrients, temperature, pressure, pH, etc.). Specifically, a small activity product and temperatures can shift any positive standard Gibbs energy to a negative overall Gibbs energy, even if standard Gibbs energy is positive (6). As a practical matter,

activation energy barriers regulate the kinetics for energetically favorable reactions, and enzymes reduce these kinetic barriers.

Metabolic diversity on Earth has not been constant with geological time. An analysis of monophyletically conserved genes among contemporary Archaea and Bacteria lineages identified 355 candidate genes possessed by the last universal common ancestor (LUCA) (7). These genes suggest LUCA was a diazotrophic autotroph utilizing the Wood-Ljungdahl pathway as a carbon fixation pathway and harvested energy using hydrogenotrophic acetogenesis and methanogenesis. These observations led (7) to conclude LUCA presided in an anaerobic environment with geochemical conditions consistent with a hydrothermal system. Weak selection pressures on metabolism evolution during the Archaean led to rapid metabolism diversification via high rates of gene birth, horizontal gene transfer, and gene duplication events (8). A gene cluster analysis performed by (8) indicates Archaean metabolisms cycled diverse sulfur species ($H_2S$, elemental S, $S_2O_3^{2-}$, $SO_3^{2-}$, and $SO_4^{2-}$), select metals (Fe, Zn, and Mn), and $NH_3$. Presumably, sulfur cycling was primarily driven by anoxygenic photosynthesis (elemental S acting as an electron donor), sulfur reduction (elemental S to $H_2S$), and sulfur disproportionation reactions (9). Furthermore, an energy budget calculated by (9) indicates iron anoxygenic photosynthesis was likely the primary metabolism fixing carbon and generating energy during the Archaean. Subsequent oxidation of Earth's atmosphere increased inventories of oxidized elemental reservoirs and increased selection pressures on anaerobic metabolisms (10). Metabolisms evolved to utilize oxidized nitrogen species ($NO_3^-$, $NO_2^-$, $N_2O$) and metals soluble in oxic conditions (Cu, Mo, Co, and Ni) (8). Although surface Earth became more oxidized, anerobic metabolisms did not undergo extinction. Contemporary subterranean biomes

with high organic matter are largely anoxic and support anaerobic microbial metabolisms (10). These biomes contain ~90% of all microbial life by mass (11).

Ideally, discerning whether a microbe can mediate a reaction, or possesses a metabolism type, is an exercise of observing a reaction in conjunction with a microbial isolate. Elucidating preferential growth conditions for microbial communities necessitates experimentation with different growth medias, building simulated environments, and consideration of obligate ecological relationships (12). Additional challenges arise when evaluating metabolisms in co-cultures, where additional techniques, such as stable isotope probing, may be necessary to determine which members perform which metabolic processes (13). In practice, most microbial lineages still lack cultured representatives using traditional culturing techniques. Often times, environmental conditions are difficult to replicate, obligate partners are unknown *a prior*, or isolates do not grow on experimentally practical timescales (14). A recent analysis quantified the proportion of cells in the environment which lack a cultured representative (15). Using distances between 16S rRNA genes to the nearest cultured representative (based on sequence similarity alignments) for all rRNA genes in publicly available metagenomes, (15) estimates that up to 87% and 64% of cells (varying upon biome) lack a cultured representative within their respective genus and phylum, respectively.

To circumnavigate issues related to culturing, culture-independent techniques are used to identify proxies, or traits. Traits can indicate whether a microbe has a metabolism. In particular, ‘omics sequencing technologies have become mainstay approaches for identifying microbial traits. With sequencing, traits are readily linked to specific microbial lineages and infer complex

geochemical cycling. For instance, genomic sequencing has linked sulfur, nitrogen, carbon, and iron cycles via metabolic syntrophy in aquifers and estuary sediments (16, 17).

Although meta-omics sequencing technology provides an alternative to culture-based approaches for characterizing microbial metabolisms, issues persist when characterizing traits from environmental communities. Perhaps the largest shortfall from meta-omics sequencing is that the observation of a gene, transcript, or protein, does not necessarily translate into a phenotype. Often 'omics data must be integrated with assays or geochemistry to infer phenotype (18). Another prominent issue with meta-omics is the continued reliance on reference databases of known gene annotations (generated from pure cultures) for assigning annotations to unknown genes. For instance, only ~50% of genes from ocean metagenomes have annotates (19). Another issue is that assembly and binning algorithms, which are used with short read sequencing, are not 100% optimal at recapitulating genomes in the environment (20). Typically, algorithm efficacy is evaluated with respect to artificial communities (21); however, the composition of environmental communities is unknown, and so experimentally validating observed metagenome-assembled genomes retrieved during sequencing experiments remains a challenge.

Despite these benign challenges, meta-omics sequencing technologies have generated a substantial amount of trait data (22–25), and if properly applied, can be informative about microbial metabolism diversity. The four chapters presented here attempt to address some quantitative and theoretical limitations associated with measuring microbial metabolic diversity.

- Chapter One, published in the journal, mSphere, as "Quantitatively Partitioning Microbial Genomic Traits Among Taxonomic Ranks Across the Microbial Tree of Life" (26), focuses on one of the outstanding questions in microbiology: how does

phylogenetic diversity correspond to metabolic diversity (27)? Is there a correlation between phylogenetic relatedness and metabolism? The recently proposed Genome Taxonomy Database (GTDB) taxonomy is the first taxonomic system quantifying relative evolutionary divergence exclusively based solely on conserved marker gene sets (28). Metabolic diversity among GTDB taxonomic ranks, phylum through genus, was quantified with cluster of orthologous groups in genomes from RefSeq v92 (23), the JGI Integrated Microbial Genomes and Microbiomes (IMG/M) database, and uncultivated bacterial and archaeal genomes from GenBank (29).

- Chapter Two, published in the journal, mSystems, as "Theoretical and Simulation-Based Investigation of the Relationship between Sequencing Effort, Microbial Community Richness, and Diversity in Binning Metagenome-assembled Genomes" (30) , explores the relationship between sequencing effort in whole-genome shotgun metagenomic sequencing experiments, community diversity, and metagenome-assembled genomes (MAG) discovery rates. A disproportionate number of microbial lineages lack cultured representatives (15), which creates a reliance on culture-independent techniques for characterizing the presence of traits. High-throughput sequencing technological has become a mainstay in the microbial ecology literature for characterizing the genomic content from lineages lacking cultured representative, yet quantitative guidance on the necessary sequencing effort as a function of targeted genome rareness is lacking in the literature (20). A theoretical model, based on a coupon collector sampling model, is proposed and related to MAG discovery rates for sequencing experiments performed on sediment, soil, human gut, and surface ocean microbial communities.

- Chapter Three presents theory for a novel measure of microbial trait functional redundancy. ‘Omics technology accessibility and utility has increased for microbial ecology studies, leading to the observation that ecosystem functions are frequently performed simultaneously by different taxa in the same community (31). The ecological role of functional redundancy is a community is not clear; however, a metric that is comparable between different communities and with respect to time is necessary in studying functional redundancy. The proposed measure for functional redundancy builds on traditional diversity theory and defines functional redundancy as the evenness in relative trait contribution by community members.
- Chapter Four presents a porewater and bulk sediment geochemical analysis of White Oak River (WOR) estuary sediments. The WOR estuary may provide a good model system for studying microbial ecology with respect to geochemistry due to the rapid change in redox reactant zonation (32) and the global carbon sink because estuaries act a conduits for terrestrial carbon transport to marine environments (33). However, temporal variance in major geochemical reservoirs for carbon and sulfur, major sources for electron acceptors and donors in WOR sediments, have not been evaluated >20 years (34, 35). Thus, carbon and sulfur budgets were re-evaluated based on recent measurements and reaction-transport modeling based on those measurements. Geochemical changes were then related to microbial metabolisms known to exist in WOR sediments.

In summary, the chapters presented here provide a framework for identifying and comparing microbial traits. These approaches will aid in linking WOR estuary sediment geochemistry with in WOR estuary sediment microbial ecosystems.

# CHAPTER ONE

# QUANTITATIVELY PARTITIONING MICROBIAL GENOMIC TRAITS AMONG TAXONOMIC RANKS ACROSS THE MICROBIAL TREE OF LIFE

This chapter is published in mSphere by Taylor M. Royalty and Andrew D. Steen. Andrew D. Steen aided in analysis, interpretation, and conception. A version is available at https://doi.org/10.1128/mSphere.00446-19 (26).

## Abstract

Widely used microbial taxonomies, such as the NCBI taxonomy, are based on a combination of sequence homology among conserved genes and historically accepted taxonomies, which were developed based on observable traits such as morphology and physiology. A recently proposed alternative taxonomy database, the Genome Taxonomy Database (GTDB), incorporates only sequence homology of conserved genes and attempts to partition taxonomic ranks such that each rank implies the same amount of evolutionary distance, regardless of its position on the phylogenetic tree. This provides the first opportunity to completely separate taxonomy from traits and therefore to quantify how taxonomic rank corresponds to traits across the microbial tree of life. We quantified the relative abundances of clusters of orthologous group functional categories (COG-FCs) as a proxy for traits within the lineages of 13,735 cultured and uncultured microbial lineages from a custom-curated genome database. On average, 41.4% of the variation in COG-FC relative abundance is explained by taxonomic rank, with domain, phylum, class, order, family, and genus explaining, on average, 3.2%, 14.6%, 4.1%, 9.2%, 4.8%, and 5.5% of the variance, respectively ($p<0.001$ for all). To our knowledge, this is the first work to quantify the variance in metabolic potential contributed by individual taxonomic ranks. A qualitative comparison between the COG-FC relative abundances and genus-level phylogenies, generated from published concatenated protein sequence alignments, further supports the idea that metabolic potential is taxonomically coherent at higher

taxonomic ranks. The quantitative analyses presented here characterize the integral relationship between diversification of microbial lineages and the metabolisms which they host.

## Importance

Recently, there has been great progress in defining a complete taxonomy of bacteria and archaea, which has been enabled by improvements in DNA sequencing technology and new bioinformatic techniques. A new, algorithmically defined microbial tree of life describes those linkages, relying solely on genetic data, which raises the issue of how microbial traits relate to taxonomy. Here, we adopted cluster of orthologous group functional categories as a scheme to describe the genomic contents of microbes, a method that can be applied to any microbial lineage for which genomes are available. This simple approach allows quantitative comparisons between microbial genomes with different gene compositions from across the microbial tree of life. Our observations demonstrate statistically significant patterns in cluster of orthologous group functional categories at taxonomic levels that span the range from domain to genus.

## Introduction

The relationship between microbial taxonomy and function is a longstanding problem in microbiology (27, 36, 37). Prior to the identification of the 16S rRNA gene as a taxonomic marker, microbial phylogenetic relationships were defined by traits such as morphology, behavior, and metabolic capacity. Cheap DNA sequencing has provided the ability to fortify those phenotype-based taxonomies with quantitative determinations of differences between marker genes, but canonical taxonomies such as the NCBI taxonomy continue to “reflect the current consensus in the systematic literature,” which ultimately derives from trait-based

taxonomies (38). Recently, (28) formalized the genome taxonomy database (GTDB), a phylogeny in which taxonomic ranks are defined by "relative evolutionary divergence" in order to create taxonomic ranks that have uniform evolutionary meaning across the microbial tree of life (28). This approach removes phenotype or traits entirely from taxonomic assignment as evolutionary distance is calculated from the alignment of 120 and 122 concatenated, universal proteins found in all bacterial and archaeal lineages, respectively. An investigation of the relationship between traits and phylogeny has not been possible until the recent publication of a microbial tree of life that is based solely on evolutionary distance. Thus, we ask the question: to what extent does GTDB phylogeny predict microbial traits?

Comparing phenotypic characteristics of microorganisms across the tree of life is not currently possible, because most organisms and lineages currently lack cultured representatives (15, 39). We therefore used the abundance of different clusters of orthologous groups (COGs) in microbial genomes, a proxy for phenotype which is available for all microorganisms for which genomes are available. Clusters of orthologous groups (COGs) are a classification scheme that defines protein domains based on groups of proteins sharing high sequence homology (40). More than ~5,700 COGs have been identified to date. COGs are placed into one of 25 metabolic functional categories (COG-FCs), which represents a generalized metabolic function (e.g., "Lipid Transport and Metabolism" or "Chromatin Structure and Dynamics"). Our analyses quantify the degree to which taxonomic rank (genus through domain) predicts the COG-FC content of genomes, and illustrate which lineages are relatively enriched or depleted in specific COG-FCs. These analyses constitute a step towards better understanding how evolutionary processes influence the distribution of metabolic traits across taxonomy as well as being able to

probabilistically predict the metabolic or functional similarity of microbes given their taxonomic classification.

## Results

The genomes analyzed in this work were compiled from a variety of different sources, including RefSeq v92, the JGI Integrated Microbial Genomes and Microbiomes (IMG/M) database, and GenBank, in order to include genomes created using diverse sequencing and assembly techniques. The integration of the RefSeq v92, JGI IMG/M, and GenBank databases resulted in a total of 119,852 genomes within the custom-curated database. Raw data, GTDB taxonomy, and associated accession numbers are provided in at https://zenodo.org/record/3361565. Among these genomes, we included only those that satisfied a set of criteria designed to ensure that each genus contained enough genomes to allow statistically robust analysis. This resulted in a set of 13,735 lineages, representing 22 bacterial phyla and 4 archaeal phyla, 67% of which have been grown in culture (Table 1.1). Most predicted open reading frames for most lineages could be assigned to a COG-FC. Across all phyla, an average of 84.3% $\pm$ 7.8% of open reading frames were assigned to a COG-FC (Figure 1.1). Genomes of the same phylum tended to group together in an initial principal component analysis (PCA) of raw COG-FC abundance (Figure 1.2A). Since this analysis was based on absolute abundance of COG-FCs in genomes, rather than relative abundance, we hypothesized that the relationship between COG-FC abundance and phylum was largely a consequence of genome size, which is phylogenetically conserved (41). Consistent with this possibility, position on PC 1 correlated closely with genome size ($r^2$=0.88; Figure 1.2B). We therefore normalized each COG-FC abundance, for each genome, to a prediction of COG-FC abundance as a function

of genome size derived from a generalized additive model (GAM; Figure 1.3; summary statistics in Table 1.2). Each GAM model was statistically significant ($p$<0.001), and all but five COG-FCs had deviance explained (analogous to adjusted $r^2$) of more than 50%. We interpret analyses of these genome size-normalized datasets as reflecting the enrichment or depletion of COG-FC abundance, relative to that expected for a given genome size, and thus, are defined as COG-FC relative abundances. PCA of these COG-FC relative abundances showed that species-level lineages still tended to group by phylum, even though the inter-phyla gradients in genome size were no longer apparent (Figure 1.2B, C). Note that attempts were made to normalize by genome size alone; however, these attempts failed to properly remove the influence of genome size. We hypothesize this was due to the nonlinear response in COG-FC abundances as a function of genome size.

To quantify the degree to which taxonomic rank explains the distribution of COG-FC relative abundances among individual genomes, we performed a permutation multivariate ANOVA (PERMANOVA) using the following taxonomic ranks: domain, phylum, class, order, family, and genus, as well as culture-status (cultured versus uncultured lineage) (Figure 1.4). The rank of species was excluded from the analysis as every lineage was unique, and thus, species would explain 100% of the data. Every rank significantly influenced the distribution of COG-FC relative abundance ($p < 0.001$), but the fraction of variance that each rank explained differed substantially: phylum explained the most variance (14.6%), followed by order (9.2%), genus (5.5%), family (4.8%), and class (4.1%). Domain explained only 3.1% of variance in COG-FC relative abundance, the least of any taxonomic rank. Culture-status was a significant correlate of COG-FC abundance ($p < 0.001$) but had virtually no explanatory power, with variance explained

<0.001%. This observation is consistent with no particular COG-FC relative abundance being systematically higher or lower in uncultured microbes relative to cultured microbes.

The variability in COG-FC relative abundances across different phyla was explored in addition to mean COG-FC composition for individual phyla (Figure 1.5). The evolutionary distance in COG-FC content was measured for all lineages in respect to the phylum COG-FC centroid (Figure 1.5A). The variations in calculated distances for all lineages within a given phylum were compared across the entire phylum (Figure 1.5A). Among all phyla, the *Crenarchaeota* differed the most from the phylum centroid, indicating the largest amount of genomic variation in terms of COG-FC content, followed by *Patescibacteria* and *Cyanobacterota*. The least variable phyla were the *Synergistota*, *Marinisomatota*, and *Fibrobacterota* phyla, respectively (Figure 1.5A). We explored the possibility that the amount of variance in lineages from the phylum centroid was a function of the number of lineages in the phylum. In other words, did the COG-FC content of some genomes seem less variable simply because they had been undersampled? A plot of the average distance of lineages from their phylum's centroid (i.e., the center of the mass of all genomes in the trait space) versus the number of lineages in the phylum reveals that increased sampling caused an apparent increase in the variability of traits within a phylum. This increase in variability across the phylum began to resemble an asymptotic line after approximately 100 genomes were sampled (Figure 1.5B). We modeled the data using both a saturating model (equation 1.1) and a linear model to test this observation. The saturating model described the relationship substantially better than a linear regression, as determined by the Akaike information criterion (AIC; $\Delta$AIC=10.5). Coefficient A of the saturating model, which represents the value of the asymptote, was estimated to be

0.75±0.15 ($p$<0.001). Coefficient B, which represents how quickly the function approaches the asymptote, was 0.43±0.30 ($p$=0.17). Coefficient C, representing an offset employed to address the fact that all the log-transformed distances have negative values, was -1.63±0.14 ($P$<0.001). This means that observing approximately 100 lineages in a phylum is sufficient to assess the variance in trait space representing half of all potential variance for that phylum (0.13). Note that this represents the effect of incorporation of the shift parameter, coefficient C.

We sought a qualitative sense of how the distribution of COG-FC relative abundance related to phylogeny. To achieve this, we quantified the average COG-FC relative abundance for each COG-FC in each genus These values were then visualized on a genus-level phylogenetic tree (Figure 1.6) utilizing concatenated ribosomal protein sequences published by (28). Several notable features appear in COG-FC relative abundance at the phylum level. For example, among the four Archaeal phyla represented here, *Thermoplasmatota* appears unique, with high COG-FC relative abundances in cell motility and depletion in every other category. In general, the COG-FC content of bacterial lineages appeared more variable than the archaeal lineages at all taxonomic resolutions. The clade consisting of *Bacteroidota*, *Spirochaetota*, and *Verrucomicrobiota* were notably depleted in the less-variable COG-FCs, including energy production and conversion, amino acid transport and metabolism, and carbohydrate transport and metabolism, among others. Another prominent feature is the near-ubiquitous elevation in COG-FC relative abundance of cell motility, secondary metabolites biosynthesis, transport, and catabolism, lipid transport and metabolism, and intracellular trafficking, secretion, and vesicular transport COGs in Proteobacteria. A notable dichotomy in the COG-FC relative abundance of RNA processing and modification within the *Proteobacteria* mirrors the division of the two

largest clades within the proteobacteria. Overall, relative abundance data qualitatively appears consistent with phylogenetic relationships, albeit, occurring on different taxonomic levels.

The relationship between individual COG-FC relative abundances and taxonomic ranks was appeared largely variable (Figure 1.6). For instance, most variation in RNA Processing and Modification occurred at higher taxonomic ranks such as phylum and class while Secondary Metabolites Biosynthesis, Transport, and Catabolism varied at lower ranks such as order. To quantify this relationship, we applied a variance components model to proportion the variance explained by different taxonomic ranks (Figure 1.7). Domain and culture-status was excluded from this analysis as variance explained becomes imprecise when a factor has less than 5 groups (42). Consistent with the PERMANOVA results (Figure 1.4), COG-FC relative abundances were best explained by the taxonomic rank, phylum. In contrast to the PERMANOVA, the taxonomic rank, class, appeared to have reasonable explanatory power for a select set of COG-FCs. In general, the overall explanatory power for taxonomic rank appears to decrease at lower taxonomic ranks.

Lastly, to gain a sense of "notable" COG-FCs associated with different phyla, we calculated the mean COG-FC across all lineages in a given phyla and compared these values against the $85^{th}$ and $15^{th}$ percentiles for all lineages in our custom-curated database. All COG-FCs which were significantly ($p<0.05$; based on a $10^5$-iteration bootstrap analysis) greater or less than the $85^{th}$ and $15^{th}$ percentiles, respectively, are shown in Table 1.3. Each archaeal phylum was enriched or depleted three-to-nine COG-FCs, whereas most bacterial phyla were enriched or depleted in in three to four COG-FCs. A few exceptions arose, such as *Fibrobacterota* was deplete in eight COG-FCs, *Nitrospirota* A was enriched in four and depleted in five, and

*Proteobacteria* was the only phylum not heavily enriched or depleted in any COG-FCs.

## Discussion

We observed that the abundance of COG-FCs within individual lineages tentatively grouped according to phyla after variable reduction via PCA (data not shown). Furthermore, PCA scores along PC1 correlated strongly with genome size ($r^2$=0.88; Figure 1.2A). It has been observed that genome size, and thus, number of genes, is a phylogenetically conserved within microbial clades (41). The conserved nature of genome size across phylogeny implied that phylogenic groupings may be an artifact of genome size. Thus, the normalization of COG-FC abundances by genome size to properly characterize the relationship between COG-FC and phylogeny. We performed the normalization using the slope from a GAM regression which modeled COG-FC abundance as a function of genome size. The COG-FC normalization removed the influence of genome size ($r^2$=0; Figure 1.2B) while retaining phylogenic groupings (Figure 1.2C).

The PERMANOVA (Figure 1.4) and analysis of diversity of genomic composition within phyla (Figure 1.5) showed that microbial lineages exhibit characteristic relative abundances of COG-FC, and that the extent of variation varies among taxonomic ranks. Of all the taxonomic ranks, phylum was the most powerful predictor of COG-FC relative abundances, which is consistent with observations that phylum can be informative of microbial function (e.g., 24–26). Lower taxonomic ranks such as genus and family had approximately half the explanatory power of the taxonomic rank, phylum. Many studies focus on metabolic coherence of individual traits and regularly find traits conserved on the family level (36, 37, 46). The discrepancy between

previous observations and our observation likely relates to how we characterize patterns in metabolic potential. These studies characterize trait function based on phenotype observation, protein structures, and pathway components. Such characterizations are effective metrics for characterizing finer units of taxonomy, such as genus, but do not scale to coarser units of taxonomy, such as phylum. In contrast, COG-FCs provide a coarse metabolic description which scales with coarser units of taxonomy (46). The tradeoff of the approach used here is that, by analyzing COG-FCs, we lose information about specific genes or potential metabolic functions but gain the ability to apply a consistent analysis across an entire genome and across the entire microbial tree of life. Thus, the extent that observed patterns (Figure 1.2) reflect phenotypically-expressed differences among lineages is unknown. Nonetheless, the statistical robustness of the relationship between all taxonomic ranks and COG-FC patterns suggests that evolutionary processes (e.g., horizontal gene transfer, vertical gene transfer, duplications, deletions, etc.) control the preponderance of different COG-FCs across lineages.

The role that individual evolutionary processes play in influencing COG-FC relative abundances at a given taxonomic rank is likely variable. For instance, horizontal gene transfer is more common among more closely related lineages (47) and thus, likely promotes increased levels of similarity at lower taxonomic ranks. At higher taxonomic ranks, vertical processes may be more important. The asymptote in the mean $\log_{10}$-distance from the centroid as function of lineages in a phylum suggests that identifying more lineages for more poorly represented lineages should expand the diversity of COG-FCs that are found, whereas phyla that were adequately sampled (at least ~1000 lineages) exhibited comparable variability in COG-FC distributions (Figure 1.5B). Since many more than ~1000 distinct lineages of each phylum are

likely to exist (48), we propose that the taxonomic rank of phylum implies a fairly consistent degree of diversity in COG-FC distribution. To the extent that phenotype matches genotype at the level of COG-FC distributions, therefore, we expect that typical phyla exhibit similar phenotypic diversity. A notable exception is the phylum *Crenarchaeota,* which were far more diverse than would be expected based on the number of lineages sampled. The *Crenarchaeota*, as defined in the GTDB, collapsed members of several phyla that had been designated separately under previous taxonomies, including lineages that had previously been assigned as *Crenarchaeota, Thaumarchaeota*, *Euryarchaeota*, *Verstraetearchaeota*, *Korarchaeota*, and *Bathyarchaeota* (28). It is possible that the relationship between marker genes used in the GTDB and the rest of the genome is unusual for this clade, compared to other phyla, or that the GTDB classification of *Crenarchaeota* is lacking in some other way.

Although the ranks, genus and family, explained relatively little of the variance in COG-FC distribution, examples of consistent colored blocks were evident at every taxonomic resolution in Figure 1.6, indicating higher or lower relative abundances of specific COG-FCs were conserved across each taxonomic rank in some parts of the phylogenetic tree. This is explained by 'distantly' (i.e., non-sister clades) related clades occupying similar COG-FC trait-space. Our variance components model accounted for the hierarchical nature of taxonomic lineage by partitioning the explanatory power that individual taxonomic ranks had for individual COG-FC relative abundances (Figure 1.7). Consistent with Figure 1.5, different COG-FCs appeared most controlled at different taxonomic ranks. For instance, the COG-FC, Coenzyme Transport and Metabolism, was almost entirely explained by the taxonomic rank, phylum. This observation is consistent with previous assessments suggesting that enzyme cofactors are deeply

conserved at the phylum level (49, 50). Similarly, the COG-FC, Carbohydrate Transport and Metabolism, was best explained the taxonomic ranks, genus and family, which is consistent with previous observations that large amounts of variability exist for hydrolase traits at lower taxonomic ranks (27, 36, 37). Ultimately, the variability in explanatory power on COG-FCs by different taxonomic ranks supports the notion that evolutionary processes operate on microbial metabolisms at different timescales depending on which component of the metabolism is in question.

The coherence in metabolic potential at higher taxonomic ranks may help explain the distribution of microbial clades across ecological niches. Analyses of habitat associations (27, 41, 50) found phylum-level patterns in lineages occupying niches which supports the idea that there is a relationship between higher taxonomic ranks, metabolism, and niche. Our analysis provides quantitative evidence to this idea by demonstrating coherence in metabolic potential with broad-scale patterns in genomic data. The question remains: how well do the observed COG-FC relative abundances reflect expressed functional traits (i.e., phenotype) across these lineages? It is difficult to address this question systematically, but some of the relative abundances and depletions here appear consistent with known physiologies of clades. For instance, *Rickettsiales* were depleted in nucleotide metabolism and transport, consistent with previously observed lack of a metabolic pathway for purine synthesis among five example *Rickettsiales* (51). Another example is the depletion in the COG-FCs for energy production and conversion, amino acid transport and metabolism, and carbohydrate transport and metabolism within the *Bacteroidetes*, *Spirochaetes*, and *Chlamydiales* clade. This clade is known to contain many host-dependent pathogens and symbionts (52–54), which are often depleted in these COG-

FCs (55).

The GTDB classification is the first fully algorithmic and quantitatively self-consistent microbial taxonomy that can be applied across the tree of life (28). By standardizing the meaning of taxonomic ranks, it creates an objective basis on which to compare microbial functionality to phylogeny. The analyses presented here demonstrate compositional patterns exist for genomic traits which can be explained by different taxonomic ranks. Furthermore, the proportion of variance explained for individual COG-FCs was partitioned as a function of taxonomic ranks. These quantitative relationships elude to the idea that evolutionary processes operate on different timescales for different components of microbial metabolisms and supports previous notions that a relationship exists between higher taxonomic ranks, metabolism, and ecological niches.

## Materials and Methods

### Genome Database Curation

All bacterial and archaeal genomes from the RefSeq database v92 (23), uncultured bacterial and archaeal (UBA) metagenome-assembled genomes (MAGs) reported in (28, 29), bacterial and archaeal MAGs from Integrated Microbial Genomes and Microbiomes (IMG/M), and bacterial and archaeal single amplified genomes (SAGs) from IMG/M were curated into a single database. All genomic content within the curated database is referred to as “genome(s)” for simplicity. Genomes were assigned taxonomy consistent with the Genome Taxonomy Database (GTDB) using the GTDB toolkit (GTDB-Tk) v0.2.1 (28). The GTDB-Tk taxonomic assignments were consistent with reference package GTDB r86. Lineages which did not receive a genus classification, due to the absence of a reference lineage were excluded from analyses. In

total, 6.1% of the total number of genomes from the initial database met this condition. Due to bias in the abundance of strains in specific clades (e.g., *E. coli*), the lowest taxonomic rank considered during our analysis was species. The COG-FC relative abundances (see below) were averaged together for all strains within a given species. An exception was made for lineages which shared a genus classification but lacked a species classification. In this scenario, each genome was treated as an independent lineage. In total, 10.9% of the total number genomes analyzed (i.e., had a genus assignment) met this condition. Lastly, only genomes belonging to genera with at least ten unique species in the database were retained. This criterion ensured enough data to generate meaningful statistics during our PERMANOVA. The final database is summarized in Table 1.1. The genus-level phylogenetic tree was generated from concatenated protein sequence alignments published in (28).

**COG Functional Category Identification, Enumeration, and Normalization**

Genes were predicted from individual genomes and translated into protein sequences using Prodigal v.2.6.3 (56). The resulting protein sequences were analyzed for COGs (40). COG position-specific scoring matrices (PSSMs) were downloaded from NCBI's Conserved Domain Database (27 March 2017 definitions). COG PSSMs were BLASTed against protein sequences with the Reverse Position Specific-BLAST (RPS-BLAST) algorithm (57). Following previously a reported protocol (57), we used an E-value cutoff of 0.01 to assign COGs with RPS-BLAST. The retrieved COGs were assigned to their respective COG functional categories (COG-FCs; 25 in total) and the abundance of each functional category was tabulated using cdd2cog (58) for each genome. The abundance for individual COG-FCs was normalized by the respective COG-FC standard deviation across all lineages. For the COG-FCs, extracellular structures and nuclear

structures, the standard deviation was 0. Consequently, data could not be normalized, and thus, these two categories were discarded from all analyses.

COG-FC abundances were normalized by their respective regression slopes of COG-FC abundance for a given genome as a function of genome size. COG-FC abundances were modelled as a function of genome size for individual categories using a generalized additive model (GAM) with a smoothing term due to the pairwise response to genome size (Figure 1.3). We used the gam function from the R package, mgcv (59). In some instances, regression fits were visibly skewed by high-leverage data points. High-leverage data were filtered using the influence.gam function in the mgcv package. Data in the 99.5% percentile for influence were excluded when performing regression analysis but were included in all downstream analyses. All regressions were significant with $p<0.001$.

**Principal Component Analysis (PCA)**

We performed PCA on the normalized COG-FC abundances and relative abundances. Prior to PCA, assumptions of normality were achieved by performing a boxcox transformation on individual COG-FC abundance and relative abundance distributions with the boxcox function from the R Package, MASS (60). The resulting distributions were then scaled by the respective COG-FC standard deviation calculated from all genomes. PCA was performed using the princomp function from the R package, stats (61).

**Quantifying COG-FC Variance Explained by Taxonomic Rank**

We performed permutational multivariate analysis of variance (PERMANOVA) using the adonis function from the R package, vegan (62). The taxonomic ranks domain, phylum,

class, order, family, and genus as well as cultured-status were used as test categorical variables for quantifying variance in COG-FC relative abundance explained by the mean taxonomic rank centroids. The default, 999 permutations test, was performed using each categorical variable. Distances were calculated between mean phyla COG-FC relative abundance centroids and the respective genomes within that phyla by performing an analysis of multivariate homogeneity of groups dispersions with the betadisper function from the R package, vegan (62). The distance matrix used for both the adonis and betadisper analyses was generated calculating Euclidean distance on the normalized COG-FC relative abundance.

The mean $\log_{10}$-distance from phylum centroid for each phylum and modeled with the following equation, which represents a hyperbola shifted on the x-axis to ensure that mean distance is zero when n = 1:

$$log_{10}(MeanDistance) = \frac{A(log_{10}(n)-1)}{B+log_{10}(n)-1} + C \quad (1.1$$

where *A*, *B*, and *C* are fit coefficients and *n* is the total number of lineages in the given phylum. The Akaike Information Criterion was calculated with the fit from eq 1 using the AIC function from the R package, stats (61).

A variance component model was performed using the lme function from the R package, nlme (63). The proportion of variance explained by the taxonomic ranks, phylum, class, order, family, and genus, was determined for each individual COG-FC. Domain and culture-status were not evaluated due to imprecise results generated from factors that only have 2 groups (42). Lineage was treated as a random intercept, where individual taxonomic ranks were nested within one another in a hierarchical manner (R notation: ~1|phylum/class/order/family/genus).

Confidence intervals were determined by performing a 500-iteration bootstrap analysis with the variance component model. During the bootstrap analysis, genomes were randomly sampled with replacement.

# Appendix One

## Tables

**Table 1.1.** A summary of the custom-curated genome database used in this work.

| Unique Domains | Unique Phyla | Unique Classes | Unique Orders | Unique Families | Unique Genera | Unique Lineages | Cultured Lineages | Uncultured Lineages |
|---|---|---|---|---|---|---|---|---|
| Bacteria | *Actinobacteriota* | 3 | 9 | 22 | 50 | 2286 | 2115 | 171 |
| | *Bacteroidota* | 3 | 7 | 19 | 50 | 1606 | 741 | 865 |
| | *Campylobacterota* | 1 | 1 | 6 | 8 | 270 | 203 | 67 |
| | *Cyanobacteriota* | 2 | 3 | 4 | 7 | 119 | 84 | 35 |
| | *Deinococcota* | 1 | 1 | 2 | 2 | 44 | 44 | 0 |
| | *Desulfobacterota* | 2 | 2 | 2 | 4 | 43 | 23 | 20 |
| | *Elusimicrobiota* | 1 | 1 | 1 | 1 | 22 | 0 | 22 |
| | *Fibrobacterota* | 1 | 1 | 1 | 1 | 34 | 22 | 12 |
| | *Firmicutes* | 3 | 10 | 23 | 48 | 1543 | 1356 | 187 |
| | *Firmicutes A* | 2 | 7 | 10 | 31 | 600 | 304 | 296 |
| | *Firmicutes B* | 1 | 1 | 1 | 1 | 22 | 11 | 11 |
| | *Firmicutes C* | 1 | 2 | 3 | 4 | 53 | 32 | 21 |
| | *Fusobacteriota* | 1 | 1 | 2 | 2 | 40 | 40 | 0 |
| | *Marinisomatota* | 1 | 1 | 1 | 1 | 10 | 0 | 10 |
| | *Nitrospirota* | 2 | 2 | 2 | 2 | 30 | 6 | 24 |
| | *Nitrospirota A* | 1 | 1 | 1 | 1 | 14 | 2 | 12 |
| | *Patescibacteria* | 6 | 16 | 26 | 36 | 707 | 0 | 707 |
| | *Proteobacteria* | 3 | 25 | 59 | 163 | 5589 | 3952 | 1637 |
| | *Spirochaetota* | 3 | 4 | 4 | 6 | 153 | 89 | 64 |
| | *Synergistota* | 1 | 1 | 1 | 1 | 19 | 2 | 17 |
| | *Thermotogota* | 1 | 1 | 2 | 2 | 23 | 15 | 8 |
| | *Verrucomicrobiota* | 2 | 4 | 5 | 7 | 84 | 16 | 68 |
| Archaea | *Crenarchaeota* | 1 | 1 | 1 | 2 | 68 | 7 | 61 |
| | *Euryarchaeota* | 2 | 2 | 2 | 2 | 45 | 26 | 19 |
| | *Halobacterota* | 4 | 5 | 7 | 9 | 164 | 97 | 67 |
| | *Thermoplasmatota* | 1 | 1 | 2 | 9 | 147 | 0 | 147 |
| Total | 26 | 50 | 110 | 209 | 450 | 13,735 | 9187 | 4548 |

**Table 1.2.** Fit statistics for GAM regressions modeling COG-FC abundance as a function of genome size.

| COG Functional Category | P-value | F-value | Deviance Explained |
|---|---|---|---|
| RNA Processing and Modification | * | 467.9 | 15.8 |
| Chromatin Structure and Dynamics | * | 404 | 14.2 |
| Energy Production and Conversion | * | 7145 | 74.1 |
| Cell Cycle, Control, Cell Division, and Chromosome Partitioning | * | 3378 | 57.9 |
| Amino Acid Transport and Metabolism | * | 6118 | 71.5 |
| Nucleotide Transport and Metabolism | * | 5845 | 70.5 |
| Carbohydrate Transport and Metabolism | * | 3941 | 61.8 |
| Coenzyme Transport and Metabolism | * | 11714 | 82.7 |
| Lipid Transport and Metabolism | * | 3609 | 59.7 |
| Translation Ribosomal Structure and Biogenesis | * | 6089 | 71.4 |
| Transcription | * | 13960 | 84 |
| Replication, Recombination, and Repair | * | 3229 | 56.6 |
| Cell Wall Membrane Envelope Biogenesis | * | 7347 | 75.1 |
| Cell Motility | * | 953 | 28.1 |
| Posttranslational Modification, Protein Turnover, chaperones | * | 8067 | 76.8 |
| Inorganic Ion Transport and Metabolism | * | 10606 | 81.3 |
| Secondary Metabolites Biosynthesis, Transport, and Catabolism | * | 3993 | 62.1 |
| General Function Prediction Only | * | 28676 | 92.1 |
| Function Unknown | * | 12393 | 83.6 |
| Signal Transduction Mechanisms | * | 7733 | 76 |
| Intracellular Trafficking Secretion and Vesicular Transport | * | 1313 | 35 |
| Defense Mechanisms | * | 4383 | 64.3 |
| Cytoskeleton | * | 351 | 12.5 |

* Denotes a statistically significant *p*-value ($p<0.05$)

**Table 1.3** Phylum highly enriched (>85$^{th}$ percentile) or depleted (<15$^{th}$ percentile) in COG-FCs and depletion. All reported categories are statistically significant.

| Phylum | Enriched COG-FC | Depleted COG-FC |
|---|---|---|
| *Actinobacteriota* | -- | 4,6 |
| *Bacteroidota* | -- | 11,18,22 |
| *Campylobacterota* | 4,6,10,15,19 | 8,12 |
| *Cyanobacteriota* | 19 | 12 |
| *Deinococcota* | -- | 15 |
| *Desulfobacterota* | 13 | 7 |
| *Elusimicrobiota* | 3,10,15 | 14,16 |
| *Fibrobacterota* | -- | 7,11,12,13,16,18,20,22 |
| *Firmicutes* | 9,12,21 | -- |
| *Firmicutes A* | 9 | 5,7,16,20 |
| *Firmicutes B* | 13,17,19 | -- |
| *Firmicutes C* | 19 | -- |
| *Fusobacteriota* | -- | 10,21 |
| *Marinisomatota* | -- | 12,14 |
| *Nitrospirota* | 6,10,17 | 8 |
| *Nitrospirota A* | 4,6,10,15 | 12,17,22,23 |
| *Patescibacteria* | -- | 7,11,16,19 |
| *Proteobacteria* | -- | -- |
| *Spirochaetota* | 4 | 5,16,19 |
| *Synergistota* | 3,4,11,16 | -- |
| *Thermotogota* | 3,4,8,10 | -- |
| *Verrucomicrobiota* | 1 | 10,12,14,21,23 |
| *Crenarchaeota* | 3,13,11,19,7,12,16,5,22 | 9,10,14,15,17 |
| *Euryarchaeota* | 2,3,13,18,22 | 5,7,10,14,15 |
| *Halobacterota* | 3,19,22 | 15,16 |
| *Thermoplasmatota* | 1,2,7,13 | 4,6,8,9,10,14, 15,16 |

| COG-FC | Table Key |
|---|---|
| Cytoskeleton | 1 |
| RNA Processing and Modification | 2 |
| Chromatin Structure and Dynamics | 3 |
| Cell Motility | 4 |
| Secondary Metabolites Biosynthesis, Transport, and Catabolism | 5 |
| Intracellular Trafficking, Secretion, and Vesicular Transport | 6 |
| Lipid Transport and Metabolism | 7 |
| Carbohydrate Transport and Metabolism | 8 |
| Defense Mechanism | 9 |
| Signal Transduction Mechanisms | 10 |
| Amino Acid Transport and Metabolism | 11 |
| Transcription | 12 |
| Energy Production and Conversion | 13 |
| Replication, Recombination, and Repair | 14 |
| Cell Wall/Membrane/Envelope Biogenesis | 15 |
| Inorganic Ion Transport and Metabolism | 16 |
| Cell Cycle Control, Cell Division, and Chromosome Partitioning | 17 |
| Function Unknown | 18 |
| Coenzyme Transport and Metabolism | 19 |
| Post-translational Modification, Protein Turnover, and Chaperone | 20 |
| Nucleotide Transport and Metabolism | 21 |
| General Function Prediction Only | 22 |
| Translation Ribosomal Structure and Biogenesis | 23 |

**Figures**

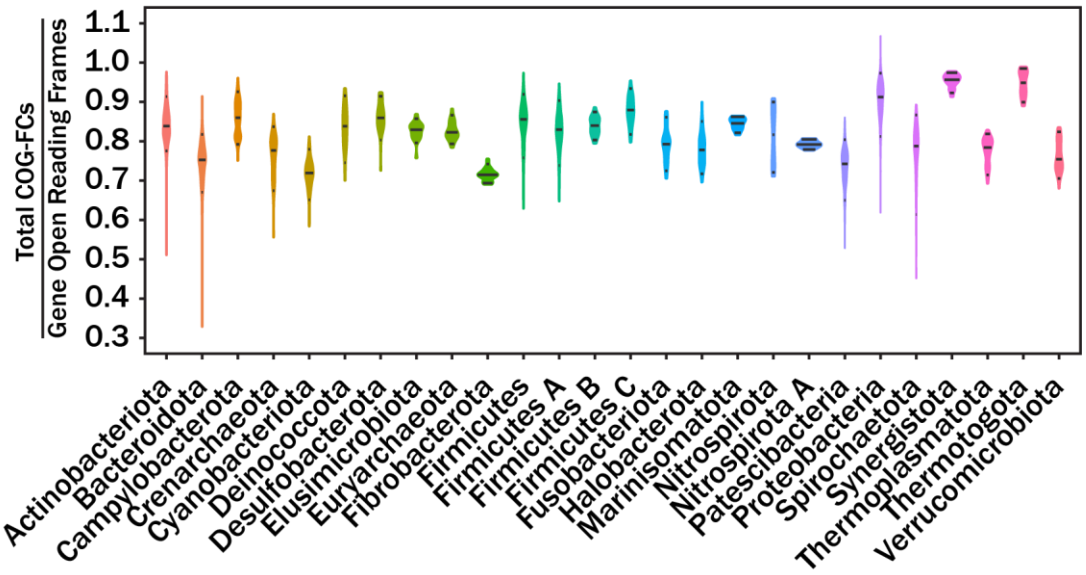


**Figure 1.1.** Violin plots showing the distribution for the ratio of total COG-FC annotations in a genome to the total number of open reading frames for each phylum.

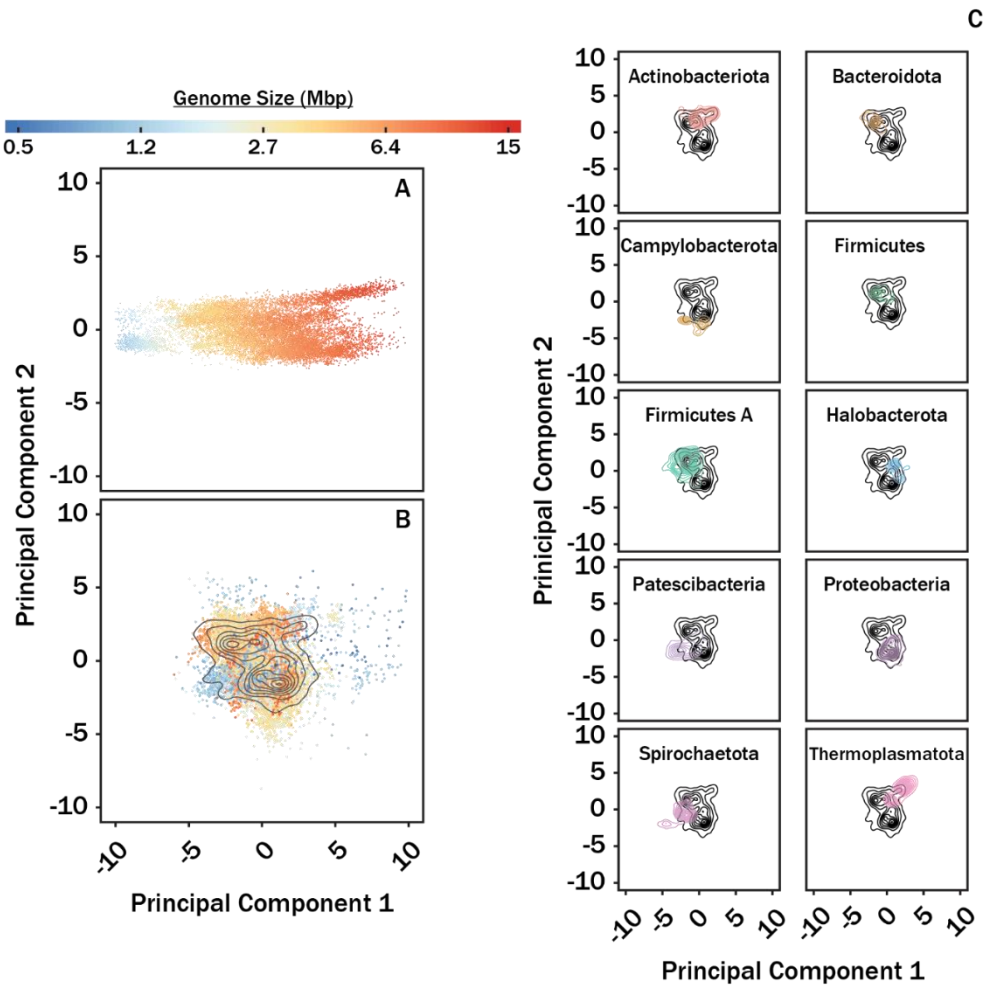


**Figure 1.2.** PCA plots of COG-FC abundances (A) and relative abundances (B and C). Individual data points are colored by genome size in panels A and B. The data presented in panel A were not normalized by genome size, while the data in panels B and C were normalized by genome size. Black contours on panels B and C correspond to density plots for all genomes shown in panel B. Colored contours in panel C correspond to the respective lineage labels. For panel A, PC1 explained 71% and PC2 explained 7.0% of the variance. For panels B and C, PC1 explained 21% and PC2 explained 16% of the variance. Panel C corresponds to only the top 10 most abundant phyla analyzed as described for Table 1.1.

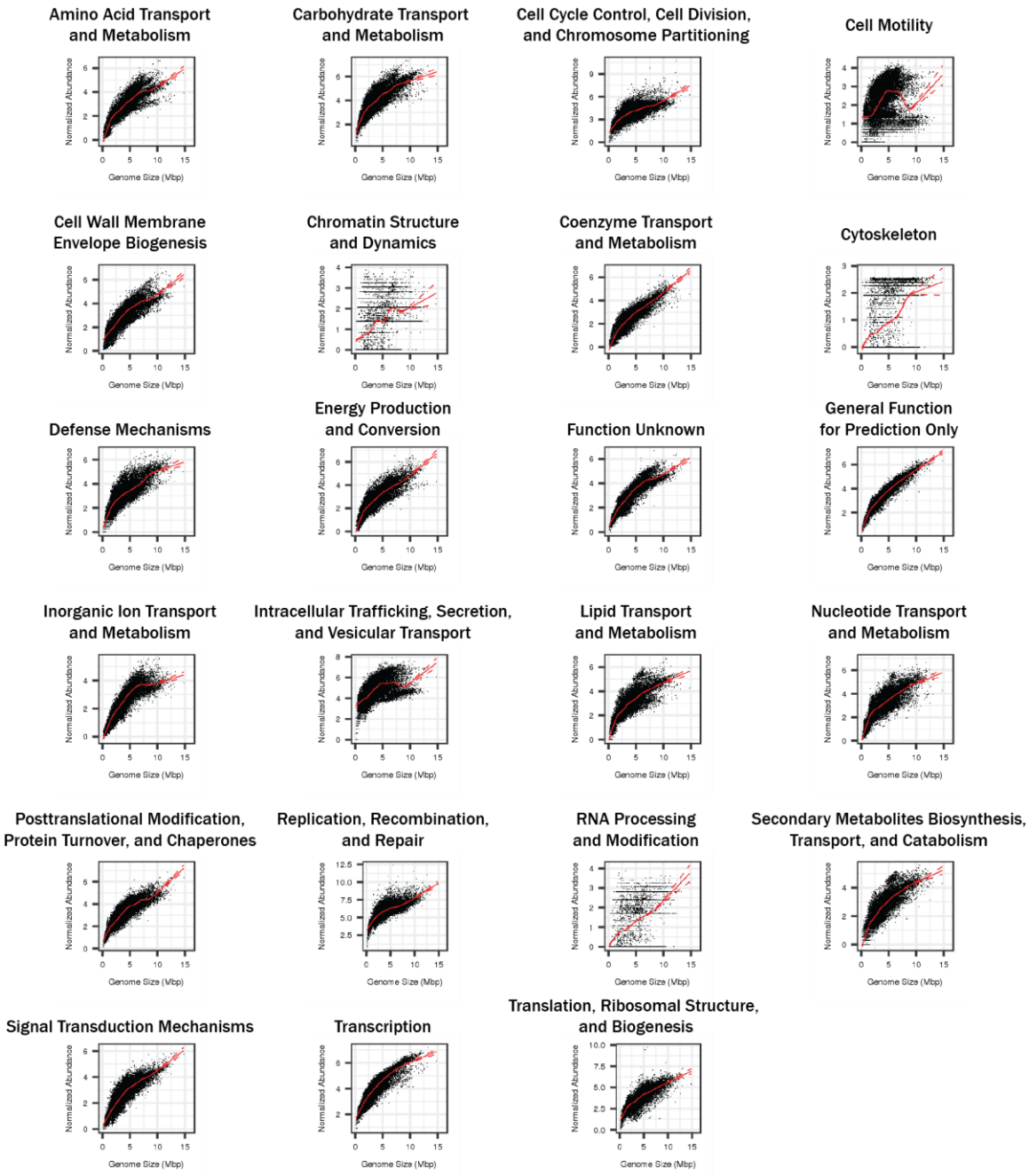


**Figure 1.3.** GAM regressions modeling COG-FC normalized abundance (standardized) as a function of genome size. Solid red lines correspond to mean fit. Upper and lower red dashed lines correspond to 95th percentile confidence intervals.

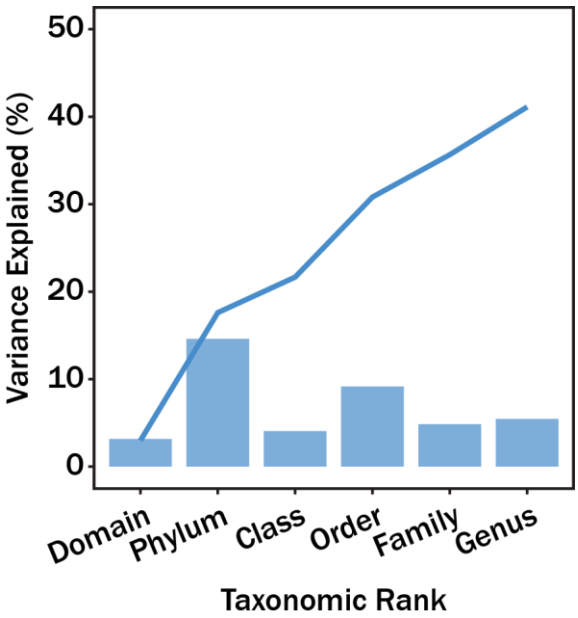


**Figure 1.4.** The average variance in COG-FC enrichment explained by different taxonomic ranks (bars) and the cumulative variance explained by taxonomic ranks (lines). All variance explained by taxonomic ranks was significant ($p$<0.001). The $F$-value for domain, phylum, class, order, family, and genus, was 726.0, 128.8, 38.76, 34.4, 11.2, and 5.1, respectively.

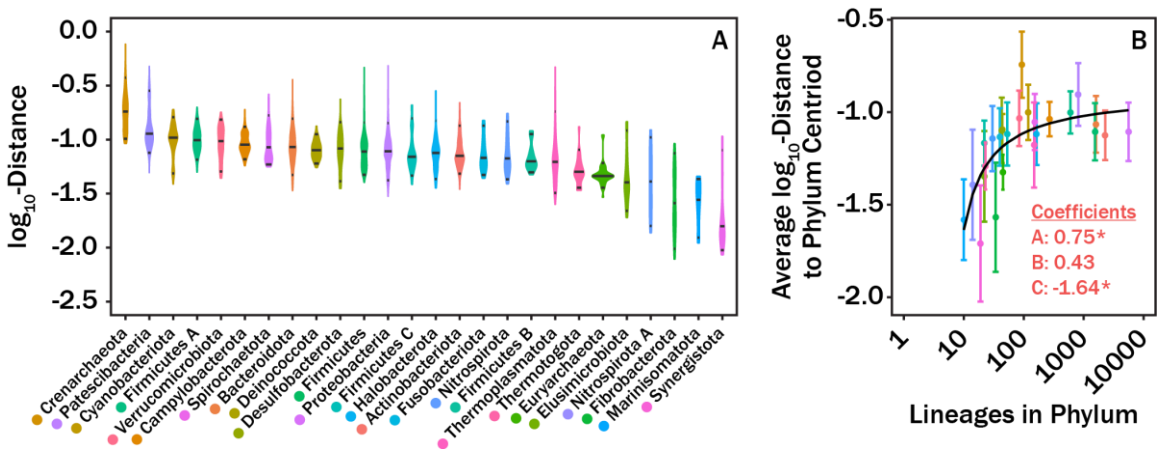


**Figure 1.5.** Violin plots showing the distribution in $\log_{10}$-distance for lineages from their respective phyla mean COG-FC centroid (A) and average $\log_{10}$-distance to phyla mean COG-FC centroid as a function of number of lineages in a given phylum (B). Coefficients in panel B correspond to fit parameters from eq 1. The * symbols denote significance as defined in the text. We note three outliers: the *Crenarhaeota* are characterized by unusually high diversity of COG-FCs distributions, and the *Synergistota* and *Fibrobacterota* are characterized by an unusually low diversity of COG-FC distributions.

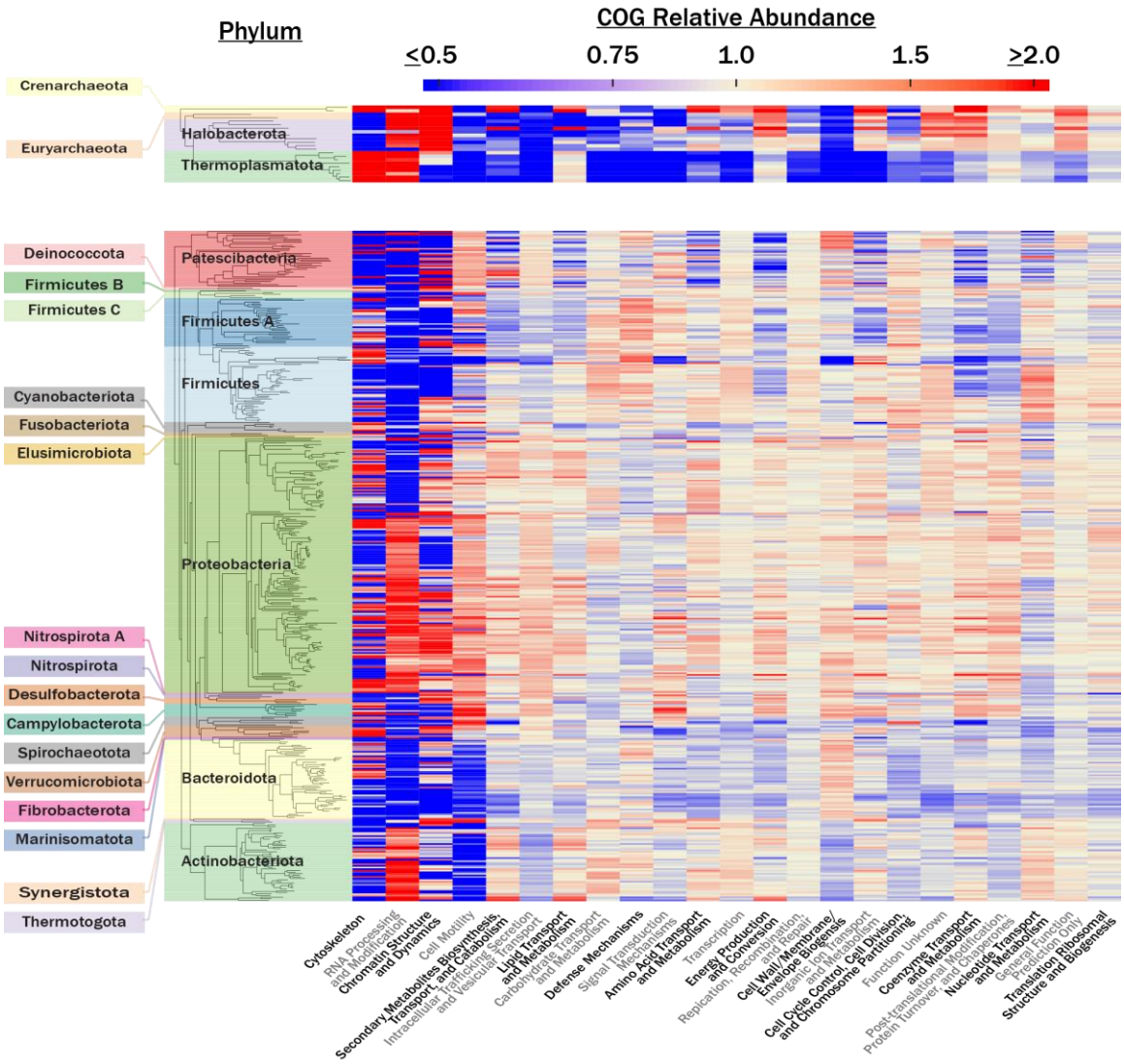


**Figure 1.6.** A heat map showing the average COG-FC enrichment for all archaeal (A) and bacterial (B) genera. Categories were arranged from left to right along the x axis in order of decreasing total variance in enrichment across all lineages. Clades were organized along the y axis using phylogenetic relatedness based on the reported concatenated protein sequence alignments in (28).

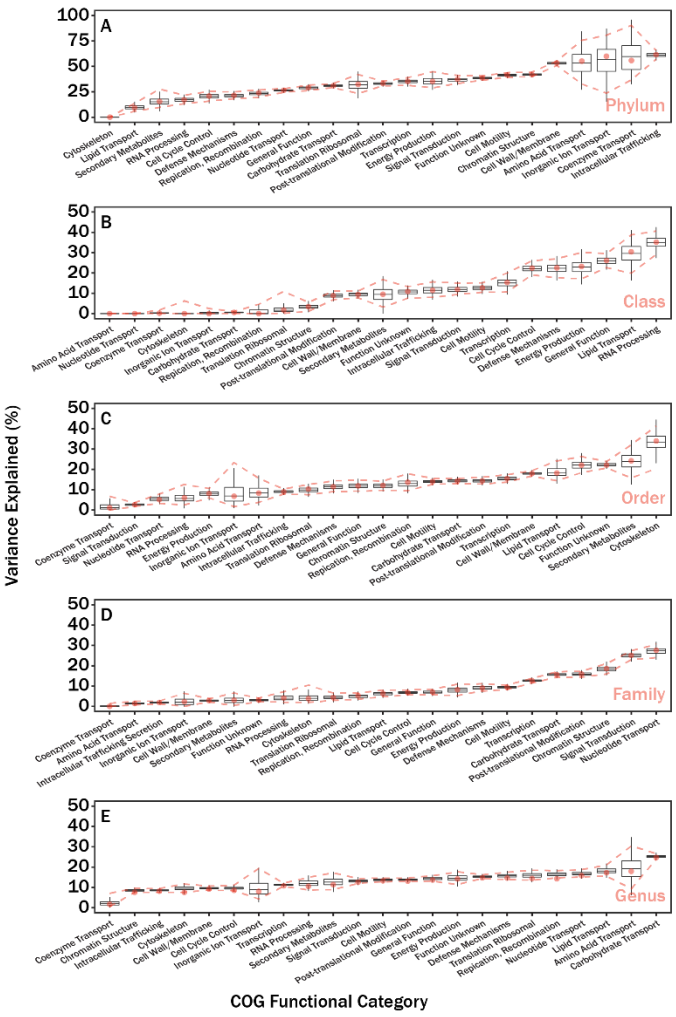


**Figure 1.7.** Results from a variance component model. Lineage was used as a nested random effect (intercept) for all COG-FCs. The proportion of variance explained is partitioned by phylum (A), class (B), order (C), family (D), and genus (E). Box plots correspond to the variability in variance explained from the bootstrap analysis, red dashed lines correspond to 95% confidence intervals calculated from the bootstrap analysis, and red circles correspond to the variance explained by analysis of all data in Table 1.1. Note that the titles for COG-FCs are shortened; the full category names are shown in Figure 1.6 and Table 1.2.

# CHAPTER TWO

# THEORETICAL AND SIMULATION-BASED INVESTIGATION OF THE RELATIONSHIP BETWEEN SEQUENCING EFFORT, MICROBIAL COMMUNITY RICHNESS, AND DIVERSITY IN BINNING METAGENOME-ASSEMBLED GENOMES

This chapter is published in mSphere by Taylor M. Royalty and Andrew D. Steen. Andrew D. Steen aided in analysis, interpretation, and conception. A version is available at https://doi.org/10.1128/mSystems.00384-19 (30).
.

## Abstract

We applied simulation-based approaches to characterize how sequencing effort influences the properties of genomes identified in metagenomes assembled from short read sequences. An initial analysis evaluated the quantity, completion, and contamination of complete-genome equivalents, a bioinformatic-pipeline normalized metric for metagenome-assembled genomes (MAGs), as a function of sequencing effort on four preexisting sequence read datasets taken from a maize soil, an estuarine sediment, the surface ocean, and the human gut. These datasets were variously subsampled to simulate the effect of sequencing effort on MAG retrieval. Modeling suggested that sequencing efforts beyond what is typical in published experiments (1 to 10 Gbp) would generate diminishing returns in terms of MAG binning. A second analysis explored the theoretical relationship between sequencing effort and the proportion of available metagenomic DNA sequenced during a sequencing experiment as a function of community richness, evenness, and genome size. Simulations from this analysis demonstrated that while community richness and evenness influenced the amount of sequencing required to sequence a community metagenome to exhaustion, the effort necessary to sequence an individual genome to a target fraction of exhaustion was only dependent on the relative abundance and genome size. The software tool, GRASE, was created to assist investigators further explore this relationship. Re-evaluation of the relationship between sequencing effort and binning success in the context of the relative abundance of genomes, as opposed to base pairs,

provides a framework to design sequencing experiments based on the relative abundance of microbes in an environment rather than arbitrary levels of sequencing effort.

## Importance

Short read sequencing with Illumina sequencing technology provides an accurate, high-throughput method for characterizing the metabolic potential of microbial communities. Short read sequences can be assembled and binned into metagenome-assembled genomes, thus shedding light on the function of microbial ecosystems that are important for health, agriculture, and Earth system processes. The work presented here provides an analytical framework to select sequencing effort as a function of experimental goals in metagenome-assembled genomes creation projects. We hope that the results presented here, as well as GRASE will be valuable to researchers planning sequencing experiments.

## Introduction

The reconstruction of high-accuracy short read sequences into metagenome-assembled genomes (MAGs) is a powerful approach to characterize microbial metabolisms within complex communities (64). The recent creation of ~8,000 MAGs from largely uncultured organisms across the tree of life (29), the spatial characterization of microbial metabolisms and ecology across Earth's oceans (65), and the characterization of the potential impact that fermentation-based microbial metabolisms have on biogeochemical cycling in subsurface sediment environments (17) provide a few examples of how MAGs have helped to elucidate the relationships between microbial ecology, microbial metabolisms, and biogeochemistry.

Sampling environmental microbial DNA involves selecting a target environment, sequencing effort, bioinformatic pipeline software and parameters, metabolism characterization software (i.e., for gene identification and similarity searches) and databases (Figure 2.1). At present, there is little information to guide how much sequencing is appropriate to achieve scientific goals in such experiments (20). This gap in knowledge is partly attributed to the unknown structure of target microbial communities. A further challenge is that the accuracy and efficiency of bioinformatic pipelines are often difficult to characterize, and thus obscure the relationship between sequencing effort and MAG retrieval. Recent estimates compiled by (20) suggest that typical metagenomic shotgun sequencing experiments usually sequence between 1 Gbp and 10 Gbp. Researchers require more precise guidance to select an appropriate shotgun sequencing effort in order to maximize information and minimize cost for their specific experimental question.

Illumina sequencing technology is currently the most popular platform to generate metagenomic shotgun sequences (20). Previous investigators established theoretical relationships between contig formation rate (66) and single genome coverage (67) as a function of short read sequencing effort. On the community level, heuristic approaches have been proposed for evaluating community-level coverage to increases in sequencing effort. For example, it has been proposed to utilize short read redundancies as a function of sequencing effort to estimate community level coverage (68). Without *a priori* knowledge of the microbial community structure, practical application of these methods to estimate MAG retrieval as a function of sequencing effort is hindered.

Here we present two distinct analyses which constrain the relationship between the quantity of Illumina metagenomic shotgun sequences and the community-level sequence coverage. First, we applied a theoretical model and numerical simulations to estimate the minimum sequencing effort needed to sequence a metagenome to a target fraction of exhaustion. Our theoretical model is unique compared to previous models (66, 67) in that we characterize sequencing effort in the context of community structure and consider all sequenceable combinations of *k*-mers in a metagenome. Second, we performed *in silico* experiments to simulate the effect of sequencing effort on retrieved MAGs for Illumina sequence datasets. Coupling results from the two analyses provides a framework for investigators to define sequencing experiments in the context of selecting a rarity and fraction of exhaustion for a desired target genome when sequencing a community. The patterns presented here can be used to guide sequencing effort decisions in future sequencing projects when MAG reconstruction is a primary goal.

## Results

### Theoretical and Numerical Sequencing Effort Simulations

Using equation 2.6, we calculated the number of sequence reads required to sequence four hypothetical microbial communities to exhaustion: a perfectly even, moderately uneven, highly uneven, and a lognormally-distributed structured community (Figure 2.2A-D). Here, we define sequencing a community to exhaustion as sequencing all possible combinations of DNA *k*-mers within a genome. Note that, to simplify the model, this model treats identical *k*-mers (i.e., same DNA sequence) at different locations in the genome as mathematically different *k*-mers. The expected number of sequence reads to fully sequence the hypothetical communities was

linear after log-transforming both expected sequences and metagenome size (unit of unique *k*-mers), suggesting a power-law relationship between metagenome size and the number of sequence reads required to sequence the metagenome to exhaustion (Figure 2.2E). For all community structures, the slope of the relationship between log-transformed sequencing effort (sequenced reads) and metagenome size (unique *k*-mers) was within 1% of 1.06. The structure of the population strongly influenced the number of reads required such that more-even community structures required far fewer reads than less-even structures.

A limitation from using equation 2.6 for modeling sequencing effort is that it only estimates the sequencing effort for sequencing a metagenome to exhaustion. We circumnavigated this limitation by applying a numerical simulation to estimate the sequencing effort necessary to sequence a metagenome to a fraction of exhaustion for the same community structures analyzed earlier. The numerical simulation was validated by comparing the sequencing effort predicted from the numerical simulation and those from equation 2.6 when sequencing a metagenome to exhaustion (Figure 2.3). Again, the relationship between metagenome size and sequencing effort fit well to a power law. This observation was independent of the selected target fraction of exhaustion. Nonetheless, an increase in target fraction of exhaustion resulted in uniform increases in sequencing effort (log units) when the community structure was fixed; however, the rate of this increase varied with community structure evenness.

We were interested in quantitatively relating community evenness to sequencing effort. These communities ranged from perfectly even ($a$=0, equation 2.8) to more uneven ($a$=0.02, Figure 2.4A). Evenness was quantified using the Pielou evenness index, which expresses Shannon diversity relative to the diversity of a perfectly even community (69). The sequencing

effort required to characterize genomes depended strongly on both the evenness and the target fraction of exhaustion (Figure 2.4B). Again, less-even communities required more sequence reads than more-even communities. The strength of this relationship also depended on the target fraction of completion. A community with Pielou evenness of 0.97 required 3 orders of magnitude more sequencing effort to sequence a metagenome to a target fraction of exhaustion than a perfectly even community while the same community only required about 42% more reads to sequence 50% of the metagenome.

The inherent limitation to the theoretical and numerical constraints presented above is that community structure is not known *a priori*. Nonetheless, a simple line of rationalization can be applied to circumnavigate this issue for practical applications of our model. Equation 2.6 constrains expected sequencing effort based on the proportion of the population under consideration. That is, we can limit the model to only consider DNA *k*-mers associated with a set of genomes that represent a certain fraction of the community. Here we limit the population to such that our model predicts the expected sequencing effort of the rarest genome that we wish to sequence. Inherently, any member more abundant than the selected rarest, will also be sequenced to the minimum coverage and depth of the selected rarest. To determine the rareness of a given genome within a metagenome, the total number of unique *k*-mers within a genome (equation 2.2) is scaled by the true abundance of the microbe and divided by the size of the metagenome (equation 2.7). Thus, the proper application of this rationale requires a desired target fraction of exhaustion, an assumed genome size, and the relative abundance of the rarest genome to sequence. Numerical simulations, like those described earlier, were performed to determine the sequencing effort necessary to a achieve a target fraction of exhaustion. The difference here is

that these simulations analyzed the sequencing effort necessary to sequence a genome of a certain rareness to a target fraction of exhaustion, whereas the simulations above analyze the effort necessary to sequence the entire community. A GAM model was built from simulation outputs to extrapolate sequencing effort required for relative abundances of less than 1% as simulations with lower abundances became computationally too intensive (GAM; Figure 2.5). The GAM shows expected sequencing effort required for microbial genome sizes 0.5, 2, 5, 10, and 20 Mbp, target genome completeness fractions from 0.5 to 1.0, and genome relative abundances from 1 to 0.0001. The smooth dimensions for target fraction, genome size, and fraction of the metagenome community were 50, 6, and 29, respectively. To normalize for different sequence read length, sequence reads were converted to bases and ranged from $1\times10^{7}$ to $1\times10^{15}$ total bases. More bases were required to sequence microorganisms 1) when the genome was relatively rarer in the community, 2) to achieve better coverage of the genome, and 3) when average genome sizes were larger.

**Rarefying MAG Binning as a function of sequencing effort**

We rarefied four sequence read datasets to nine different fractions of the complete sequence datasets in triplicate, respectively. The rarefied datasets were then assembled and binned for a total of 108 analyzed metagenomes. The sum of medium- and high-quality MAGs as a function of high-quality bases empirically fit the Gompertz equation (equation 2.12; Figure 2.6B; parameters in Table 1.1). The sum of medium- and high-quality MAGs (henceforth referred to as quality MAGs for brevity) reduces sensitivity to whether bioinformatic pipelines tend to lump contigs into fewer, more-complete MAGs versus split them into more, less-complete MAGs. This was important for our analysis due to the large number of metagenomes

(*n*=108) which required assembly and an unsupervised binning algorithm. The large number of metagenomes made manual bin curation impractical during rarefication simulations.

For each environment, the quality MAGs as a function of simulated sequencing effort followed a sigmoidal relationship. In order to make the parameters of the fit intuitive to interpret, we fit the data to the Gompertz equation as rewritten by (70) (equation 2.12). Here, $A$, $\mu$, and $\lambda$ correspond to the maximum quality MAGs assembled with the pipeline, the maximum rate which the quality MAGs form with more sequencing, and the "lag bases," or the bases which must be sequenced prior to a sufficient number of sequence reads existing to generate overlap and form contigs (66). The predicted maximum quality MAGs varied from ~9.6 in the estuary sediment community to ~19 in the maize soil community. The predicted maximum rate that the quality MAGs increased varied from ~0.9 to ~3.8 MAGs $Gbp^{-1}$. Lastly, the minimum threshold of sequencing necessary prior to seeing quality MAGs varied from ~0.6 to ~6.1 Gbp. The *Tara Oceans* dataset, where the quality decreased at sequencing effort >20 Gbp, was an exception. For the estuary, maize, and human gut datasets, the quality MAGs yield began to asymptote with increasing sequencing efforts. The *Tara Oceans* dataset followed a similar pattern at <25 Gbp. However, when the number of sequenced bases was >25Gbp, the quality MAGs decreased and became insensitive to sequencing effort.

**Constraining MAG Rarefaction Analyses to Community Structure**

Using the relationship shown in Figure 2.5, we can convert sequencing effort to an abundance (rareness) cutoff if we assume a genome size. Genome sizes for genomes in RefSeq v92 (23) have $25^{th}$, $50^{th}$, and $75^{th}$ quantiles of 2.73 Mbp, 4.30 Mbp, and 5.14 Mbp, respectively, and provide reasonable constraints for assumptions of genome size. Figure 2.7 shows quality

MAGs retrieved for a genome of a given level of abundance that is sequenced to a target fraction of completeness for the human gut, maize soil, estuarine sediment, and the surface ocean microbiomes analyzed earlier. Correlation coefficients for the regressions used to relate log-transformed sequencing effort (base pairs) to genome relative abundance were $R$=1 for all three genome sizes evaluated (1 Mbp, 5 Mbp, and 20 Mbp). Evaluation of 1 Mbp and 20 Mbp define the range of uncertainty in predicting quality MAGs as the true size of genomes is unknown. Unlike the asymptotic response of quality MAGs to sequencing effort (Figure 2.6B-E), quality MAGs increase exponentially as the abundance cutoff decreases (note that the abundance cutoff is on a log-scale). The target genome completeness fraction was held at a constant of 0.5 for all regressions and the sensitivity of quality MAGs will change with respect to the genome abundance cutoff with different values of target genome completeness fraction.

## Discussion

We sought to establish evidence-based guidelines for selecting a sequencing effort during shotgun metagenomic sequencing experiments. The model proposed here (equation 2.6) addresses this goal. Our model establishes an intrinsic relationship between a community structure and the sequencing effort necessary to sequence members of different rareness, or relative abundance in a community (equation 2.1; Figure 2.8). It is important to emphasize that the proposed model treats individual *k*-mers within a genome as members of the total population of DNA in an environment (equation 2.4; Figure 2.8). Thus, the relative abundance of any given *k*-mer in a population is equal to the abundance of the host genome divided by the total number of *k*-mers in the entire metagenome (equation 2.7). Summing the probabilities of sequencing all individual *k*-mers form the same genome (equation 2.2; Figure 2.8), should equal the relative

abundance of the genome within the population of genomes. The theoretical model utilizes these individual probabilities of sampling *k*-mers to determine how much sequencing effort is required to sequence all possible *k*-mers in a community (Figure 2.2-2.3) and for a member of some rareness in a community (Figure 2.5). Practical sequencing challenges associated with strain-level microdiversity, extracellular DNA, and sequencer contamination are not problematic for the proposed model. However, these issues can still lead to problems during assembly. In particular, abundant lineages with a large amount of strain-level microdiversity can inhibit assembly. Strains could be treated as independent taxonomic units and ultimately, sequencing effort should be selected based on a target rareness of DNA being sequenced in the sequencer. Proper measures such as removing extracellular DNA (e.g., 55) and properly removing contamination DNA are essential to avoid skewing the sequenced DNA population. Lastly, in practice, homologous DNA regions across genomes (e.g., shared genes) will be sequenced faster than unique regions. For simplicity of our model, we cannot predict the degree of homologous DNA and simply treat unique loci at DNA as a unique *k*-mer. Nonetheless, if this information is known *a priori,* the proposed model can account for homologous DNA. Equation 2.6 calculates sequencing effort for individual *k*-mer regions based on their relative abundance ($p_j$) in respective to all *k*-mers in the metagenome. In the case of homologous DNA, the relative abundances (associated with individual genomes) would be summed together.

The theoretical probability model (equation 2.6) demonstrated that the sequencing effort to sequence a metagenome to exhaustion was predictable, regardless of community structure (Figure 2.2). Furthermore, our characterization of community complexity demonstrates that less complex communities require less sequencing to sequence all available DNA (Figure 2.4). A

limitation to equation 2.6 is that it only predicts sequencing effort to sequence a metagenome to completion. In practice, assemblers do not require every unique *k*-mer to generate an accurate contig. As long as sufficient overlap exists between sequenced *k*-mers, assemblers can generate contigs. We therefore used a numerical simulation to predict the sequencing effort necessary to sequence an individual metagenome to a target fraction of exhaustion. The numerical simulation agreed with the theoretical model when predicting the sequencing effort to sequence a metagenome to exhaustion (Figure 2.3). A last analysis explored the semi-quantitative relationship between community evenness and richness and the necessary sequencing effort to achieve a target fraction of exhaustion for a metagenome (Figure 2.4).

In practice, information about a target community structure may not be available for estimating sequencing effort. The GAM built here predicts the minimum number of sequences necessary to sequence a given fraction of a target genome as a function of average genome size and the relative abundance of the target genome in the community (Figure 2.5). Even without knowledge of a target community's structure, the GAM provides a useful constraint for designing whole-genome shotgun metagenomic sequencing experiments. The size of prokaryotic genomes is fairly constrained, with 50% of the genomes in RefSeq v92 spanning from 2.74 Mbp to 5.15 Mbp (23). The limited range in genome size allows for reasonable assumptions about genomes size for prokaryotes.

We wanted to relate our model to actual sequencing experiments. As such, the subsampling analysis on existing short read datasets (individually sampled, assembled, and binned) simulated the effect of creating MAGs from datasets of different sizes and environments. The datasets analyzed here are representative of both the sequencing effort (1 to 10 Gbp) (20)

and the types of target environments that bacterial and archaeal ecologists often investigate (72). We want to emphasize that the datasets analyzed here do not necessarily reflect all the variability in characteristics of their parent communities (i.e., these do not reflect global/temporal variations of these environments). Furthermore, a wide variety of software packages are available for all steps of MAG creation pipelines, and the quantity/quality of MAGs will depend on software selection, software configuration, and sequenced environment (Figure 2.1) (20). It is best-practice to manually curate algorithmically-created MAG bins (73). Due to the large number of metagenomes which were assembled in this work (*n*=108), manual curation was an impractical approach. As with most sequencing experiments, viral/eukarotyic DNA are likely included during assembly and binning. Thus, we do not argue that the pipeline used here is objectively optimal for generating high-quality (high completeness and low contamination) MAGs. Rather, our pipeline was configured to minimize contamination (MAGs with <10%) associated with retrieved MAGs at the expense of reduced completeness. For this reason, we reported the number of quality MAGs (medium and high-quality MAGs) rather than the actual number of MAGs. Reporting quality MAGs reduces bias associated with the generalized approach used with the binning software. As such, we interpret the datasets analyzed here as reflecting general community properties (richness, abundances, and phylogeny) which are generally known to be different from one another (17, 74–76).

All communities demonstrated similar responses to rarefaction. All subsampled depths were performed in triplicate to account for possible variation in assembly and binning. As sequencing effort increased, there was an initial lag in quality MAGs followed by a rapid increase in quality MAGs, and then diminishing returns at higher sequencing efforts. Previous

investigators modeled the response of 16S RNA gene (77–79), Hill's number diversity (80), taxon-resolved abundance (81), and gene abundance (81) as a function of sequencing effort with rarefaction curves, or collector's curves. The quality MAGs as a function of sequencing effort did not match a traditional collector's curve, which lacks an initial lag. Rather, the data appear sigmoidal. We modeled the data using the Gompertz function (equation 2.12), because its parameters can be interpreted in terms of quantities that are familiar from microbial growth curves (lag time, growth rate, and maximum density) (70). The fits to the Gompertz function illustrate that there is an optimal sequencing effort for MAG creation efforts corresponding to the upper shoulder of the Gompertz curve ("late log phase"). When sequencing effort is too close to $\lambda$, MAGs bin poorly, and when sequencing effort is too great, the number and quality of MAGs per unit sequencing effort (and therefore cost) is low. We speculate that our choice of pipeline, and specifically the fact that we discarded contigs <3kb, caused poor performance at higher sequencing effort for the *Tara Oceans* dataset. Species-level microdiversity and interspecies homologous DNA can cause "bubbles" which impair assembly in larger data sets (82, 83). Improved assembly would likely have yielded more quality MAGs for our assembly of the largest subsets of the *Tara Oceans* data.

Metagenomic shotgun sequencing experimental designs should be rationally designed such that sequencing effort is selected to capture a desired fraction of a target microbial genome. Investigators should be cognizant of the rarest microbial genome desired to be characterized as well as the degree of characterization of that microbial metagenome when designing a sequencing experiment. To that end, we have built a tool, Genome Relative Abundance to Sequencing Effort (GRASE), to report estimated sequencing effort required to capture a defined

fraction of a genome as a function of the relative abundance of the corresponding microorganism in the community and average genome size. This R-based GUI app can be accessed online at http://adsteen.shinyapps.io/grase and is archived at http://github.com/adsteen/GRASE, from which it can be downloaded and run locally.

When the sequence read datasets analyzed here (Table 2.2; Figure 2.6) are re-evaluated in the context of the relative abundance of a microbial metagenome ($g_{MG}$) sequenced to a target fraction of exhaustion (0.5), quality MAGs increase appreciably in response to minor increases in deeper characterization of the community metagenome (Figure 2.7). This observation contrasts the quality MAGs response to sequencing effort (in base pairs), where substantial increases in sequencing effort (by contemporary standards) leads to diminishing returns in quality MAGs. It is important to note that abundance cutoff and sequencing effort are interchangeable; however, quality MAGs response to changes in the respective predictor (i.e., base pairs versus abundance cutoff) alters the optics of the collector's curve. Modest increases in sequencing effort contribute minor amounts to extending abundance cutoff. A substantial amount of genomic and metabolic data can be gained from targeting rarer microbes (metagenomic abundances <0.005), with the caveat that whole-genome shotgun sequencing technology (as well as computational power) requires significant increases in either the number or length of reads generated per run.

## Materials and Methods

### Defining the Microbial Metagenome and Sequencing Probability

Here we draw on set theory to provide a theoretical grounding for our *in silico* simulations described below. The expected number of sequences to sequence a fraction of an

individual microbe's genome can be modeled with probability theory by defining a community metagenome with set theory. Figure 2.8A-E provides a cartoon illustrating the application of this set theory on a hypothetical microbial population, *G*. *G* contains unique genomes (*g*) with finite abundances (*n*). The definition of microbial species is somewhat contentious (84). Here we define *g* as a genome which is unique in length and composition for all loci when compared to all other genomes in a community. As such, the species richness (*s*; unique *g*) of *G* will vary on how it is defined and should be consistent with the objectives of the investigator. In the example community (Figure 2.8A-E), *s*=6 and the total *n*=13. Thus, *G* is represented as (Figure 2.8A):

$$G = \{n_1 g_1, n_2 g_2 \dots n_s g_s | n \in N\} \quad (2.1$$

where *s* is the species richness. When characterizing *G* via shotgun metagenomics, the $i^{th}$ genome, $g_i$, can be sequenced at *K* unique sections given a characteristic read length, *k*, and average genome size, *l*, in number of base pairs (Figure 2.8B). Thus, the number of possible *k*-mers, *K*, associated with the $i^{th}$ genome, $g_i$, within *G* is equal to:

$$K_{g_i} = l(g_i) - k + 1 \quad (2.2$$

Note that equation 2.2 considers homologous DNA as unique *k*-mers. This is for model simplification. From equation 2.2, the metagenome, $g_{MG}$, for $g_i$ is defined as the set of all *k*-mers (Figure 2.8C) or:

$$g_{MG,i} = \{g_{i,1,1+k}, g_{i,2,2+k} \cdots, g_{i,Kg_i,Kg_i+k}\} \quad (2.3$$

where the subscripts for $g_i$ represent a given *k*-sized read spanning from an arbitrary starting base pair to the arbitrary starting base pair plus *k*. By substituting $g_{MG,i}$ into all *g* for equation 2.1 (Figure 2.8D), the metagenome for a microbial community, $G_{MG}$, is derived to be:

$$G_{MG} = \{n_1 g_{MG,1}, n_2 g_{MG,2} \dots n_s g_{MG,s} | n \in N\} \quad (2.4$$

while the population of *k*-mers in the metagenome, $G_{MG}$ (Figure 2.8E), is represented as:

$$K_{MG} = \{g_{MG,1}, g_{MG,2} \cdots g_{MG,s}\} \quad (2.5$$

From equation 2.5, one can determine the cardinality, or the total number, of *k*-mers associated with $G_{MG}$ (expressed as $|K_{MG}|$). To an effect, $|K_{MG}|$ is analogous with "metagenomic richness" of an environment. When attempting to fully sequence $G_{MG}$ using shotgun metagenomics, we assume that sampling events (sequence reads) are independent and are sampled with replacement (85).

The probability of sequencing all elements in $G_{MG}$ reduces to a coupon collectors problem (86) by making the above assumptions. Using the general functional form for calculating expected samples for sampling all unique elements in a set (equation 13b in (86)), one can predict the number of sequences necessary to sequence all elements in $K_{MG}$, such that the expected number of sequences, $E(G_{MG})$, is:

$$E(G_{MG}) = \int_0^\infty \left(1 - \prod_{j \in K_{MG}} (1 - e^{-p_j t})\right) dt \quad (2.6$$

where *j* is a given element within $K_{MG}$, *t* is the number of sampling events, and $p_j$ is equal to the proportion of the $j^{th}$ *k*-sized read within a given population of *k*-sized reads. $p_j$ can be expressed as follows:

$$p_j = \frac{n_i \times j \in K_{MG}}{|G_{MG}|} \quad (2.7$$

where $n_i$ is the respective abundance for the species whose MAG contains the $j^{th}$ *k*-sized read within $K_{MG}$, and $|G_{MG}|$ is the cardinality of $G_{MG}$, or the total number of *k*-sized reads in the metagenome, $G_{MG}$.

**Modeling Expected Sequences**

Equation 2.6 provides an estimate for the total number of sequences to sequence all $K_{MG}$.

The influence of increasing species richness (i.e., $s$ in equation 2.1) on the expected number of sequences was tested for four hypothetical communities. The first community had an even structure such that all $k$-mers were equally distributed across all $K_{MG}$. In the second community, 90% of the $k$-mers were equally distributed in 50% of $K_{MG}$, and the remaining 10% of the $k$-mers were distributed equally across the remaining 50% of $K_{MG}$. This community represented a community with relatively moderate species evenness. In the third community, 90% of the $k$-mers were equally distributed across 10% of $K_{MG}$, and the remaining 10% of the $k$-mers were distributed equally across the remaining 90% of $K_{MG}$. This community represented a community with relatively low species evenness. The last community had 10 equally-sized groups. The abundance of the $k$-mers in each group was based on the function form of a lognormal community (87) which has been observed in microbial populations (e.g., (80, 88)), such that:

$$S(R) = S_0 e^{-a^2 R^2} \quad (2.8$$

where $S_0$ was treated as the maximum relative of abundance ($S_0 = 1$), $a$ was the inverse width of the distribution, $R$ was treated as the positive octave range spanning 0 to 9, and $S(R)$ represented the abundance for a given group. For the lognormal abundance distribution in Figure 2.2D, $a$ was set to a value of 0.2. Each hypothetical community started with $|K_{MG}| = 1 \times 10^6$. $|K_{MG}|$ incrementally increased at 10 equally-spaced, linear steps to a maximum of $|K_{MG}| = 1 \times 10^8$. As $|K_{MG}|$ increased, all community structures remained constant. Graphical representation of rank abundance in Figure 2.2A-D was normalized by a given $|K_{MG}|$ to reflect that populations retained the same structure even as population size varied. We defined a normalized rank abundance $r_n$ such that

$$r_n = \frac{r}{s} \quad (2.9$$

where $r$ and $s$ are untransformed rank abundance and richness, respectively. For each community, at each step, the expected number of sequences was calculated using equation 2.6. The expected number of sequences as a function of $|K_{MG}|$ were modeled with linear regressions.

Equation 2.6 gives the expected number of sequences required to sequence any sized community to exhaustion. Numerical sequencing simulations were performed to determine the number of sequences necessary to sequence a subset of all $k$-mers ($K_{\mathrm{MG}}$). These numerical sequencing simulations were applied to four hypothetical community structures described above. Numerical simulations were performed such that $|K_{MG}| = 3 \times 10^7, 4 \times 10^7, 5 \times 10^7, 7 \times 10^7, 9 \times 10^7$, and $1 \times 10^8$. During each of these simulations, the parameters read length ($k$) and average genome size ($l$) were set to 100 and $1 \times 10^6$, respectively, for all $g$. Random elements from $K_{\mathrm{MG}}$ were selected with replacement to simulate sequencing events. Numerical simulations were performed until the percentage of $|K_{MG}|$ sequenced was 50%, 70%, 90%, 95%, 99%, or 100%. A weight distribution was applied to elements in a given $K_{\mathrm{MG}}$. The weight distribution biased sequencing to reflect the relative abundances of the four hypothetical communities described above. The percentage of $|K_{MG}|$ sequenced was evaluated every $1 \times 10^7$ sequences. Numerical simulations were performed in triplicate for all $|K_{MG}|$ and all target fractions of $|K_{MG}|$.

We explored the influence of community evenness on required sequencing effort by performing sequencing simulations on 6 different lognormally-distributed communities. The numerical sequencing simulations followed the simulations described above. The 6 lognormal communities were modeled such that each community had $S_0$ =1, $R$=10, and $|K_{MG}| = 1 \times 10^7$. Values of $a$ for the 6 lognormal distributions were $a$=0, $a$=0.005, $a$=0.008, $a$=0.01, $a$=0.015, and $a$=0.02. Evenness was represented using the Pielou evenness index (69), or the ratio of the

Shannon diversity index (89) for an observed community to an even community of equal richness. Shannon diversity was calculated in the context of a metagenomes such that:

$$H_{MG} = \sum_{j \in K_{MG}} -p_j log(p_j) \quad (2.10$$

where $p_j$ is the proportion that the $j^{th}$ $k$-sized read represents among all $k$-mersin the metagenome. Thus, the Pielou evenness index (69) was calculated such that:

$$J = \frac{H_{MG}'}{H_{MG,max}} \quad (2.11$$

where $J$ was the Pielou evenness index, $H_{MG}'$ was the metagenome Shannon diversity index, and $H_{MG,max}$ represented the metagenome Shannon diversity index when all $p_j$ were equal ($a$=0).

Lastly, numerical simulations were performed to determine the sequencing effort necessary to achieve a target fraction for an individual metagenome ($g_{MG}$). Target fractions increased from 0.5 to 1 at 100 linearly-spaced intervals. The fraction of the metagenome community ($G_{MG}$) that $g_{MG}$ represented varied from 1% to 100% in 30 lognormally-spaced intervals. The target genome sizes varied such that $l$=0.5×10$^6$, $l$=1×10$^6$, $l$=2×10$^6$, $l$=3×10$^6$, $l$=5×10$^6$, $l$=10×10$^6$, $l$=15×10$^6$, and $l$=20×10$^6$. The sequencing effort for a given combination of target fraction, genome size, and fraction of the metagenome community was modeled using the GAM function (mgcv R package; (59)).

**Sequence Data Sources**

All sequence data were downloaded from NCBI's Sequence Read Archive (SRA) using the SRA Toolkit (fastq-dump –split-files) (90). Exact duplicate reads for both forward and reverse reads were removed using PRINSEQ (-derep 1; v0.20.4) (91). All sequencing datasets were limited to Illumina shotgun metagenomic paired-end reads. Four datasets were analyzed for this analysis. The first dataset was from oceanic surface water collected at 5m depth in the

Caribbean Sea as a part of the *Tara Oceans* expedition (74). The second dataset was from sediment from a depth of 8-10 cm below the surface (sulfate-rich zone) and collected at the White Oak River Estuary, Station H, North Carolina, USA (17). The third dataset was collected from maize soil (75). The last dataset was collected from human fecal samples and represented a human gut microbiome (76).

All datasets analyzed in this study are summarized in Table 2.1.

**MAG Assembly Pipeline**

The pipeline developed here followed similar pipelines described by other authors (65, 92). All sequence datasets were analyzed as follows. Trimmomatic (v0.36) (93) removed adapters and trimmed low-quality bases from individual reads. Read leading and trailing quality scores were required to be >3. The sliding window was set to 4 base pairs and filtered base pair windows with a mean score <15. Quality controlled reads were assembled into contigs using MEGAHIT (v1.1.2; --presets meta-large) (94). Due to computational limitations, assembled contigs <3000 bp long were excluded from the analysis. Redundant contigs were removed using CD-HIT (v4.6.8; cd-hi-est -c 0.99 -n 10) (95). The quality-controlled reads (i.e., after using Trimmomatic) were mapped to the remaining contigs using Bowtie 2 (v2.3.3) (96) to generate a coverage score for individual contigs. Resultant contigs were iteratively clustered into MAGs using the unsupervised, clustering algorithm Binsanity (v0.2.6) (92). Similar to (65), six initial clustering iterations were performed with the parameter, preference (-p), set to -10 (iteration 1), -5 (iteration 2), -3 (iteration 3-6). Between iterations, a refinement step (Binsanity-refine) was performed on the putative MAGs with constant preference (-p) of -25. The refined putative MAGs were evaluated for contamination and completeness using the software CheckM (v1.0.6)

(97), which uses HMMER (v3.1) and Prodigal (v2.6.3) (56). For this work, we used the recommended lineage-specific marker sets. The lineage-specific workflow defines marker genes as those appearing in >97% of a single lineage. Contigs associated with putative MAGs meeting one of the following criteria: 1) had a completeness > 90% and contamination < 10%, 2) had a completeness > 80% and contamination < 5%, or 3) had a completeness > 50% and contamination < 5% were treated as algorithmically-optimized MAGS. All other MAGs were considered low-quality MAGs. MAGs defined as high-quality were not modified any further. Contigs associated with the high-quality MAGs were not used in the subsequent reclustering and refinement steps. The contigs associated with low-quality MAGs were pooled together and reclustered during the next iteration of Binsanity clustering. After the sixth iteration, the remaining MAGs which did not fall into one of the three categories underwent additional refinement using Binsanity-refine. During this step, MAGs were iteratively refined with preference set to -10 (iteration 1), -3 (iteration 2), and -1 (iteration 3). Between each refinement step, metrics of contamination and completeness were evaluated using CheckM. Again, MAGs meeting the criteria of a high-quality category described above were not further modified. The respective contigs associated with putative MAGs were not used in proceeding refinement steps. After the last iteration of refinement, all MAGs were reevaluated for completeness, contamination, and a taxonomic rank using CheckM.

**Subsampling Sequence Read Datasets**

The effect of decreased sequencing effort was simulated by subsampling the initial sequence read datasets described above. Downloaded sequence read datasets were randomly sampled at set fractions of 1%, 10%, 20%, 40%, 60%, 80%, 90%, 95%, and 100%. Each fraction

was resampled, assembled, and binned in triplicate. Each triplicate assembly was binned independently using the MAG assembly pipeline described above.

**Modeling MAG Response to Sequencing effort**

Medium-quality and high-quality MAGs were defined based on completeness and contamination from CheckM (97). Medium quality MAGs were defined as MAGs with >50% completeness and < 10% contamination. High quality MAGs were defined as MAGs with >90% completeness and <5% contamination (73). The sum of medium and high quality MAGs as a function of sequencing effort was modeled for environmental sequence datasets using the Gompertz equation, as reformulated by (70) for use with microbial growth curves:

$$g_{eq}(A, \mu, \lambda, b) = A \times e^{-e^{\frac{\mu \times e}{A}(\lambda - b)+1}} \quad (2.12$$

where $A$, $\mu$, and $\lambda$ are fit coefficients and $b$ is high-quality bases (Gbp). MAG yield could be defined as:

$$MAG\ Yield = \frac{n_{MQ}+n_{HQ}}{A} \quad (2.13$$

where $g_{eq,max}$$n_{MQ}$ is the total medium quality MAGs derived from the subsampling experiment, $n_{HQ}$ is the total high-quality MAGs derived from the subsampling experiment, and $A$ is from equation 2.12.

**Relating MAG Response to the Theoretical Sequencing Model**

Sequencing effort (in base pairs) used for predicting quality MAGs for the sequence read datasets maize soil, an estuarine sediment, the surface ocean, and the human gut, were related to a genome relative abundance utilizing the GAM presented in the previous section. This was accomplished by setting a constant genome size and target fraction (0.5) and performing a linear regression between genome relative abundance (a genome's fraction of a community) and log-

transformed base pairs. The linear regression was performed with genome sizes of 1 Mbp, 5 Mbp, and 20 Mbp. Qaulity MAGs were predicted as a function of genome relative abundance sequenced to a target fraction of 0.5.

# Appendix Two

## Tables

**Table 2.1.** Estimates of fit coefficients for the Gompertz equation (equation 2.12) for the sum of high-quality and medium-quality MAGs as a function of sequencing depth in published data sets from ocean surface water, estuarine sediment, maize soil, and the human gut*.

| Environment | $A$ (±SE) | $\mu$ (±SE) | $\lambda$ (±SE) | Yield |
|---|---|---|---|---|
| Ocean Surface Water | 97.67 (4.15) | 5.84 (0.31) | 1.16 (0.33) | 0.88 |
| Estuary Sediment | 26.25 (2.11) | 1.63 (0.06) | 3.70 (0.24) | 0.86 |
| Maize Soil | 43.65 (1.98) | 1.43 (0.07) | 6.13 (0.60) | 0.70 |
| Human Gut | 17.49 (1.02) | 5.01 (0.46) | 0.67 (0.14) | 0.90 |

*$p$-values for all coefficients were $<<0.05$ except for gut and ocean surface water values. Yield is defined in equation 2.13 as the number of medium- and high-quality MAGs relative to the number predicted at infinite sequencing depth

**Table 2.2**. Summary of sequence datasets analyzed with the MAG pipeline.

| Environment | NCBI SRA Accession | Sequencing Platform | Total Reads* | High Quality Bases* | General Notes | Citation |
|---|---|---|---|---|---|---|
| Ocean Surface Water | ERR599029 | Illumina HiSeq 2000 | $3.372 \times 10^8$ | $3.340 \times 10^{10}$ | Caribbean Sea (5 mbsl) | (74) |
| Estuary Sediment | SRR5248164 | Illumina HiSeq 2000 | $1.137 \times 10^8$ | $1.589 \times 10^{10}$ | Sulfate Zone (8-10 cmbsf) | (17) |
| Maize Soil | SRR351473 | Illumina HiSeq 2000 | $4.727 \times 10^8$ | $3.824 \times 10^{10}$ | Surface Soil | (75) |
| Human Gut | SRR5127631 | Illumina HiSeq 2000 | $5.095 \times 10^7$ | $4.847 \times 10^9$ | -- | (76) |

* Combined forward and reverse paired-end reads.

Figures

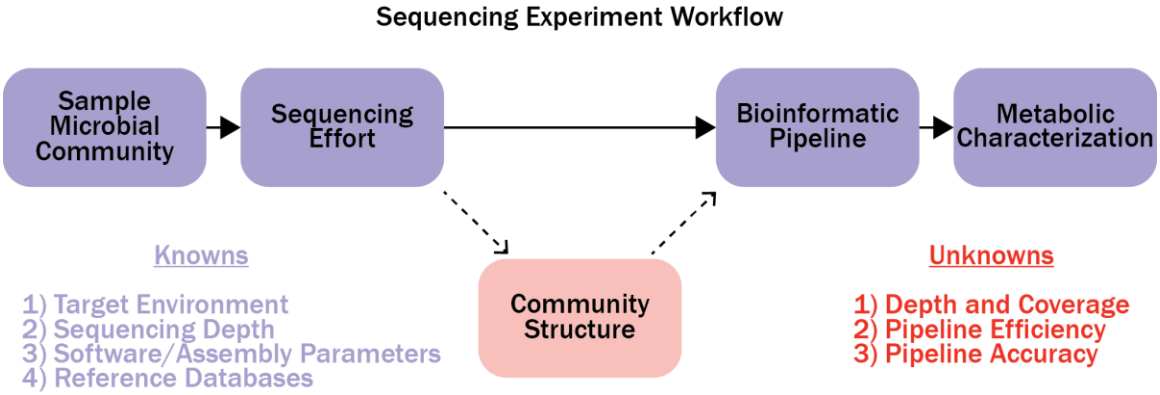


**Figure 2.1.** A flow diagram illustrating the workflow for sequencing experiments.

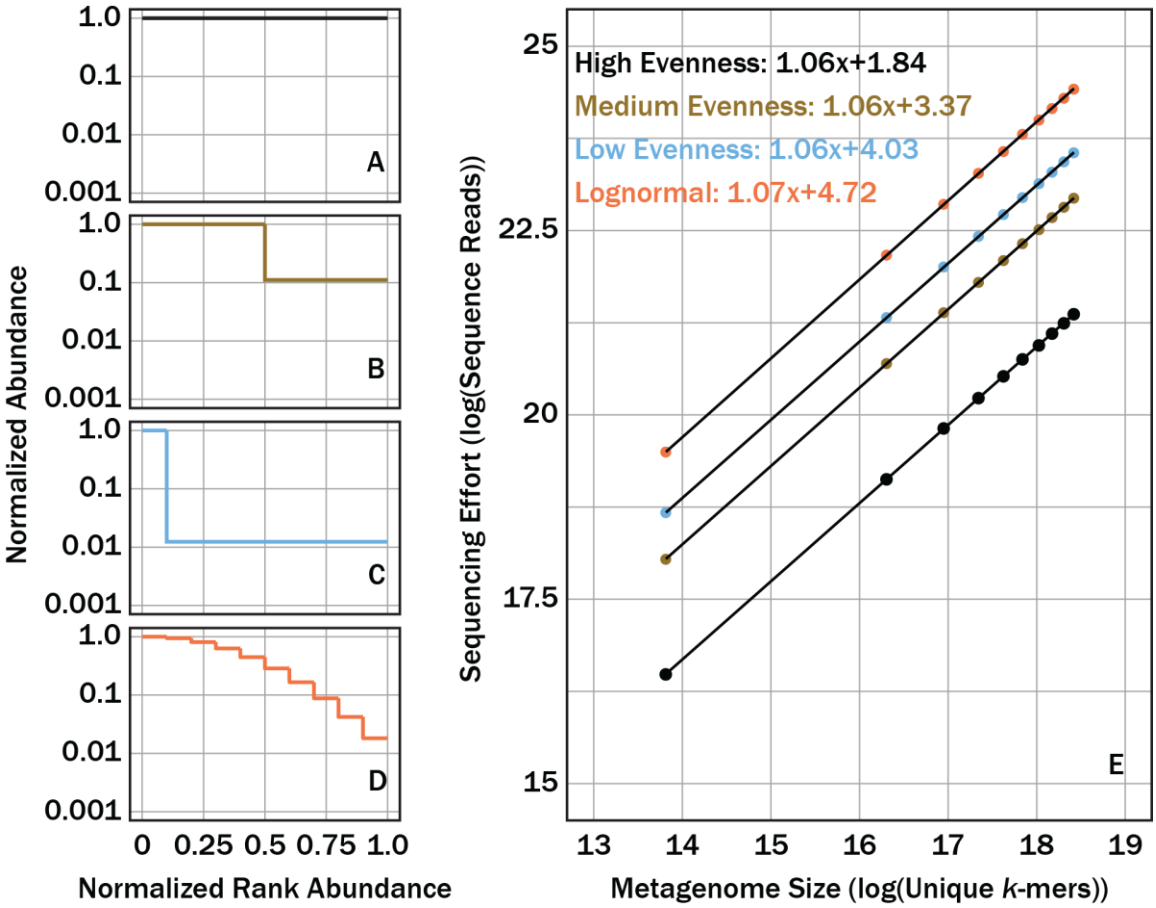


**Figure. 2.2.** Sequencing effort, with units of log of sequence reads, required to fully sequence four different community structures, one with relatively high community evenness (A), relatively moderate community evenness (B), relatively low community evenness (C), and one with a lognormal community structure (D), were predicted using linear regressions (E) and the log of metagenome size (total possible *k*-mers) as a predictor.

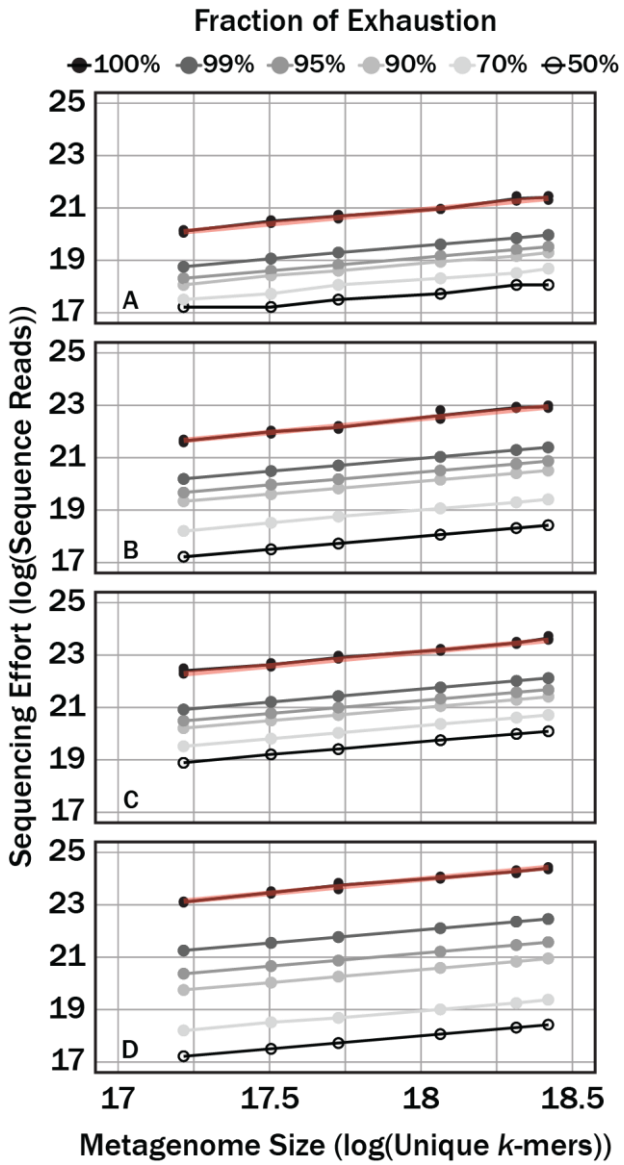


**Figure 2.3.** Sequencing effort, with units of log sequence reads, necessary to reach variable target sequencing depths (colors) for four different community structures, one with relatively high community evenness (A), relatively moderate community evenness (B), relatively low community evenness (C), and one with a lognormal community structure (D). Red translucent lines correspond with linear regression curves for the respective community in Figure 2.3E.

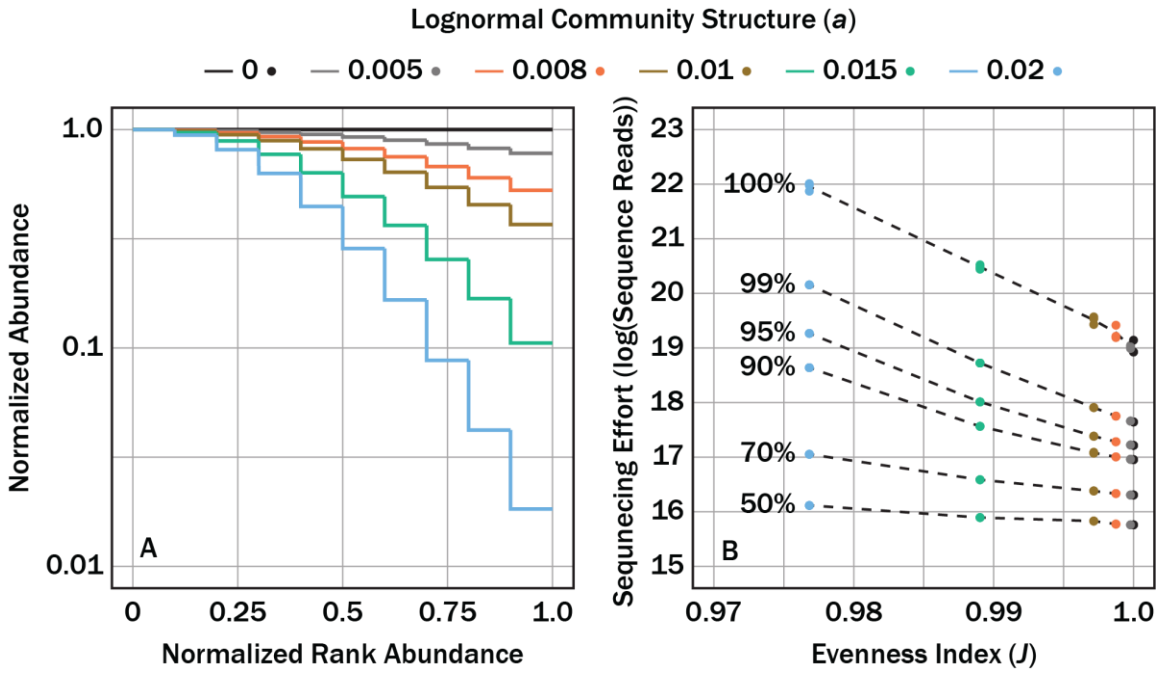


**Figure 2.4.** Numerical sequencing simulations applied to 6 hypothetical communities with different lognormal distributions that were defined by the parameter, *a*, from equation 2.7 (A). The sequencing effort, with units of log of sequence reads, necessary to sequence a target fraction of exhaustion (dashed contours) as a function of the Pielou evenness index, *J*, for a given lognormal community structure (B).

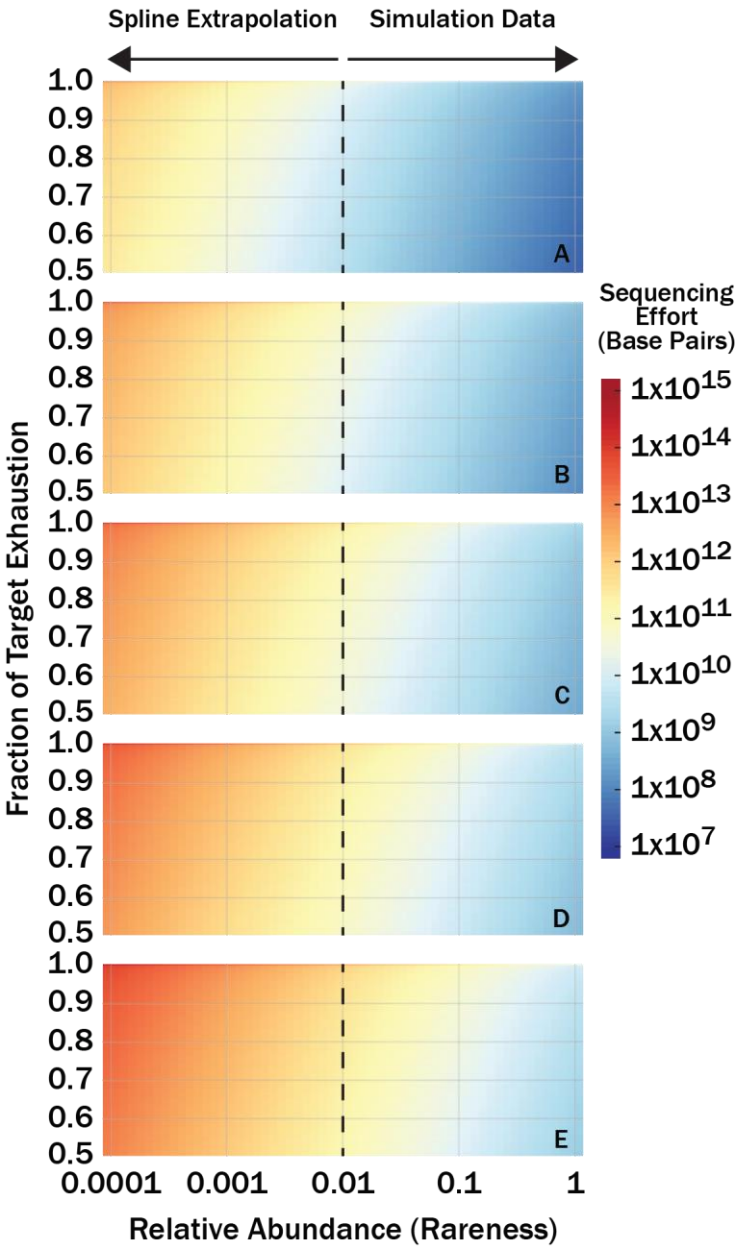


**Figure 2.5.** Numerical sequencing simulations show the number of bases (color bar) required to sequence a target fraction of a genome as a function of that genome's relative abundance in the community metagenome. Genomes evaluated were $0.5\times10^6$ (A), $2\times10^6$ (B), $5\times10^6$ (C), $10\times10^6$ (D), and $20\times10^6$ (E) base pairs long.

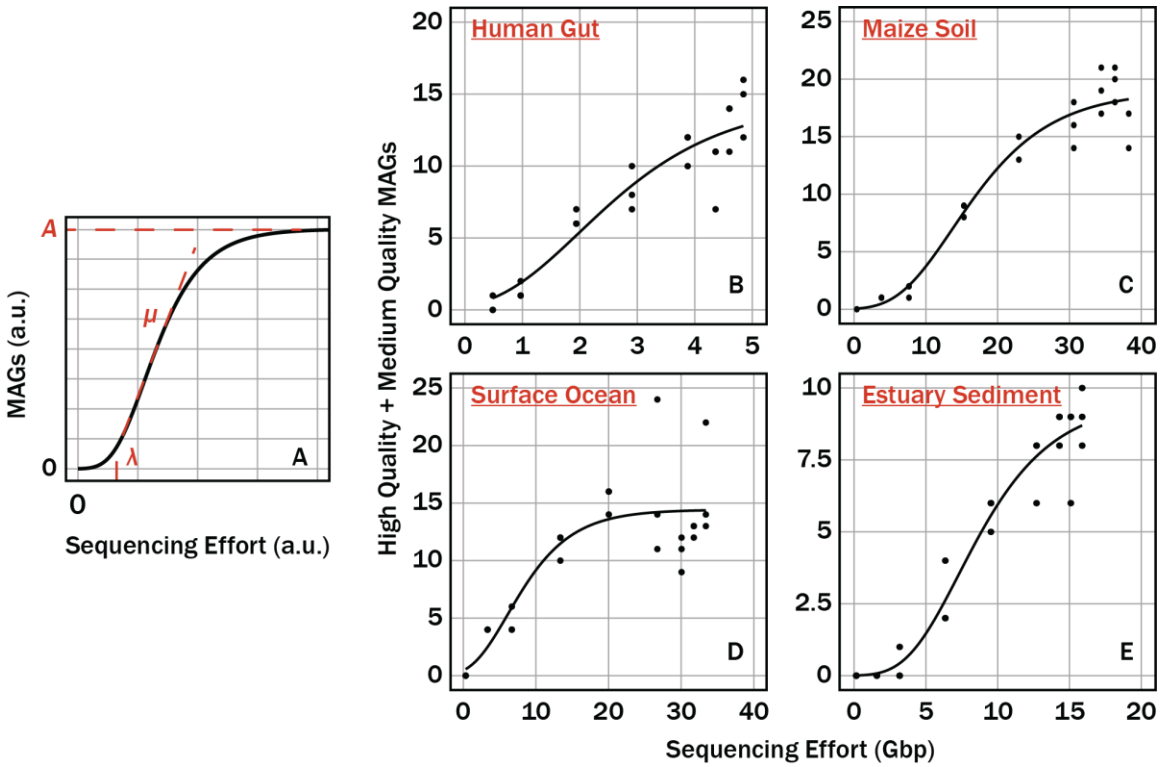


**Figure 2.6.** The influence that the parameters *A*, $\mu$, and $\lambda$ had on the Gompertz equation (A). The property of the Gompertz equation that each parameter influences is colored red. The sum of medium quality and high quality (quality MAGs) as a function of sequencing effort for the gut (B), soil (C), surface ocean (D), and sediment (E) communities. Solid lines in (B-E) correspond to nonlinear least squares fits of the Gompertz equation to the respective environmental dataset.

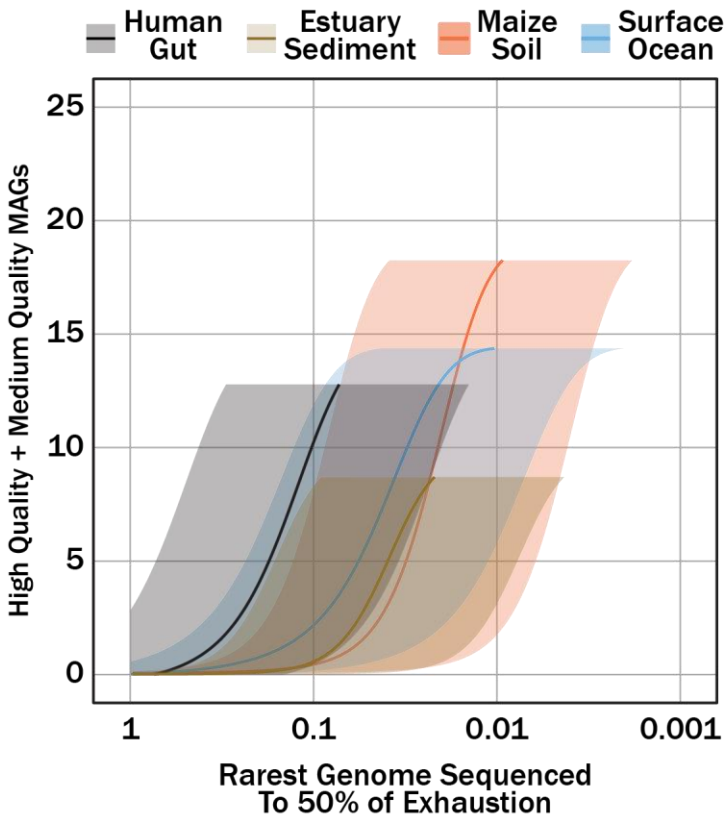


**Figure 2.7.** Quality MAGs (medium- and high-quality) as a function of rarest genome sequenced to 50% exhaustion for human gut, maize soil, estuarian sediment, and surface ocean sequence datasets. Sequencing effort was converted to genome relative abundance sequenced to target fraction using the GAM presented in Figure 2.5. The translucent shaded areas correspond to uncertainty from the target genome size (1 Mbp, or the lower-bound, to 20 Mbp, or the upper-bound) while the solid lines correspond to genome sizes 5 Mbp.

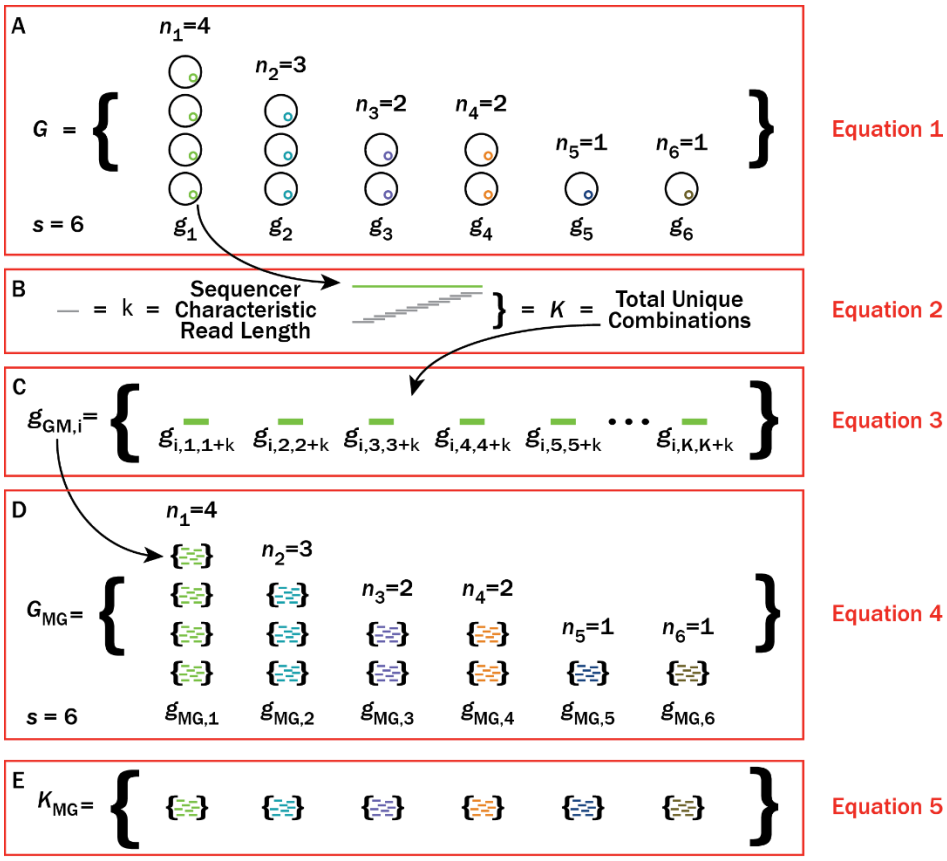


**Figure 2.8.** A cartoon illustrating an example microbial community ($G$), the metagenomes for genomes ($g_{MG,i}$) (as defined in equation 2.2) within $G$, and the overall metagenome for the given microbial community ($G_{MG}$). In this example, there are 6 genomes ($s$=6) and a total of 13 individual microbes. (A) Black circles represent individual microbes whose genomes are averaged together, $g$. The average genome, $g$, are indicated by different color inner-circles. (B) Individual average genomes can be sequenced at $K$ unique positions depending on the characteristic read length, $k$, of a sequencer. (C) All unique positions that can be sequenced for a given genome, $g$, defines the metagenome, $g_{MG}$, for the $i^{th}$ genome, $g_i$. (D) Replacing all individual genomes in (A) with metagenomes, $g_{MG}$, gives the metagenome of the microbial community, $G_{MG}$.

# CHAPTER THREE

# A QUANTITATIVE MEASURE OF FUNCTIONAL REDUNDANCY IN MICROBIAL ECOSYSTEMS

This chapter is published as a preprint at Bioxriv by Taylor M. Royalty and Andrew D. Steen. Andrew D. Steen aided in analysis, interpretation, and conception. A version is available at https://doi.org/10.1101/2020.04.22.054593 (98).

## Abstract

Although functional redundancy has received increased attention in the microbial ecology literature, no quantitative functional redundancy measurement is currently available that compares multiple communities and integrates ‘omics data rather than phenotypic traits. Here, we propose an approach for quantifying functional redundancy that uses ‘omics data. This approach, termed trait contribution evenness (TCE), is based on traditional measures of community diversity. We measure functional redundancy of a trait within a community as the evenness in relative contribution of that trait among taxa within the community. This definition has several appealing properties, including (i) TCE is an extension of established diversity theory; (ii) functional redundancy measurements from communities with different richness and relative trait contribution by taxa are easily comparable; (iii) data for a quantifiable trait (e.g., genes copies, protein abundance, transcript copies, respiration rates, etc.) are suitable for analysis. Resilience of a trait to taxa extinctions is often viewed as an ecological consequence of traits with high functional redundancy. We demonstrate that TCE functional redundancy is closely and monotonically related to the resilience of a trait to extinctions of trait-bearing taxa. Finally, to illustrate the applicability of TCE, we analyzed the functional redundancy of eight nitrogen-transforming pathways using 2,631 metagenome-assembled genomes from 47 *TARA* Oceans sites. We found that the $NH_4^+$ assimilation pathway was the most functionally redundant (0.6 to 0.7) while nitrification had the lowest functional redundancy (0 to 0.1). Here, TCE

functional redundancy addresses shortfalls of other functional redundancy measurements by providing a generalizable, quantitative, and comparable functional redundancy measurement.

## Importance

There is a necessity to integrate traditional ecological theory when quantifying community functional redundancy in microbiological studies using ‘omics data. Such an approach should allow for comparisons in functional redundancy between different samples, sites, and studies. Here, we propose measuring functional redundancy based on an expansion of already existing diversity theory. This approach measures how evenly different members in a community contribute to the overall level of a trait within a community. The approach will allow for broader evaluation of traits from ‘omics data.

## Introduction

In recent years, the emergence of ‘omics technologies has greatly improved microbial ecologists’ ability to evaluate the relationships between phylogeny, trait distributions, and environmental gradients (19, 26, 36, 99). Such studies in diverse environments have frequently found that ecosystem functions are frequently performed simultaneously by many taxa (31). This is referred to as functional redundancy (100). High functional redundancy in freshwater sediments results in stabilized porewater redox conditions (101). In surface seawater, a metagenomic analysis of glycosyl hydrolase genes demonstrates a complex successional pattern in heterotrophs that correlates with coastal phytoplankton bloom life cycles and influenced ecosystem organic carbon processing (102). In the case of human gut microbiomes, it has been

hypothesized that functional redundancy stabilizes community diversity and leads to better health (103).

A quantifiable and comparable measure for functional redundancy is a prerequisite in addressing ecological questions pertaining to functional redundancy. Concepts related to functional diversity and redundancy have been actively discussed in the macroecology literature for over 40 years (100, 104, 105). Nonetheless, there remains no universal method to quantify functional redundancy (105–108). Many of the most popular methods rely on measuring dissimilarity in traits between community members. Dissimilarities are then adjusted to consider the community member relative abundances. Most notably, the normalization of Rao's quadratic diversity by Simpson's index has been used as a quantitative measure of functional redundancy (105). In this approach, functional redundancy is quantified as the average similarity in relative trait contribution by different taxa. However, this approach does not appear sensitive to changes in concavity and convexity of community trait distributions, as demonstrated here.

In principle, methods to quantify functional diversity in macro-ecosystems should be equally applicable to microbial ecosystems. In practice, the measurement of microbial community composition and trait distribution possess unique challenges. In macro-ecosystems, tens or hundreds of species might be considered, and traits are often determined from direct field observations. In microbial ecosystems, millions of unique taxa may be present, and the appropriate level of taxonomic organization (i.e., amplicon sequence variant, operational taxonomic unit defined by a semi-arbitrary cutoff of 16S rRNA gene similarity) depends on experimental design. Traits, such as net respiration or nutrient uptake rates, are sometimes directly measurable but linking rates of specific biogeochemical transformations to specific taxa

is a challenging exercise. Most microbial cells on Earth belong to genera that lack cultured representatives, so the traits of most members of microbiomes are frequently unknown (109). Consequently, to characterize community function, the microbiological community relies on ‘omics technology of genes, transcripts, or proteins. This reality necessitates a functional redundancy measurement suitable for both phenotypic and ‘omics data.

Microbial ecologists have typically addressed these problems using two broad approaches. One common method is to compare changes in ecosystem function to changes in community composition. Stability in ecosystem function with increasing microbial diversity is often considered an empirical indication of functional redundancy (101, 110, 111). A complimentary approach is to use ‘omics data as indicators of traits (27, 112). The presence of similar traits in distinct taxa is taken as evidence of functional redundancy. This approach may stretch the definition of trait, as ‘omics data do not guarantee a phenotypic response. Nonetheless, this second approach circumnavigates the lack of taxa-specific physiological data for most environmental microbiomes. Overall, these approaches are conceptually straightforward but are largely qualitative and yield results that are challenging to compare among studies.

Here, we use diversity theory (113) to establish a quantitative approach to measure functional redundancy of microbiomes using genomic traits. We argue that functional redundancy is best understood as the evenness of the contribution of taxa to a community-wide trait. Communities are functionally redundant if taxa contribute equally to an ecosystem function (100, 110). We measured this by first defining two measures of abundance which are treated of equal importance: taxa abundance distributions and trait abundance distributions, or levels. Taxa abundance distributions correspond to conventional species abundance distributions (114). Trait

abundance distributions correspond to relative trait levels for individual taxa, i.e., the total amount of a trait in taxon. By taking the product of these two abundance distributions, we measured the relative contribution by taxa to the overall trait within a community, or the trait contribution distribution.

Traditional approaches for measuring diversity and evenness are applicable to trait contribution distributions, as the trait contribution distribution fulfills mathematical properties defining a community (113). We find trait contribution evenness (TCE) appealing because it is easily generalizable. Theoretically, any type of trait can be used (e.g., gene copies, enzyme kinetics, nutrient uptake, etc.) if the trait can be assigned to taxa and the relative abundance of those respective taxa are known. We demonstrate the efficacy of TCE by correlating functional redundancy against a conceptual idea commonly associated with functional redundancy, the risk of losing ecosystem function due to taxa extinction. Lastly, we analyze functional redundancy for eight nitrogen-transforming pathways found in metagenomic data (115) in *TARA* Oceans draft metagenome-assembled genomes (MAGs) (65). We hope that the formulation of functional redundancy fills a void in the microbial ecology literature by providing a generalized tool for generating higher quality measurements of an important parameter in microbial communities.

## Results and Discussion

### Representing microbial communities as taxa abundance, trait abundance, and trait contribution distributions

We represent microbial community traits using two distributions: the distribution of taxon abundance within the community, and the distribution of trait quantity among each taxon (Figure 3.1A). We define ‘trait’ liberally to include genes, transcripts, or proteins (“potential traits”).

This definition has been adopted in many studies when inferring microbially-mediated biogeochemistry and microbial ecology (1, 16, 17, 102, 116). Furthermore, traits encompass processes catalyzed by microbes such as nutrient uptake or respiration ("realized traits"). The choice of traits to analyze is always central in experimental design. However, for the purposes of the method described here, traits may represent anything that can be quantifiably apportioned among taxa, including binary presence/absence apportionments.

Because trait levels are measured as non-negative, real numbers, both taxa relative abundances (equation 3.1) and trait abundances (equation 3.2) can be represented as relative abundance distributions (Figure 3.1B). We determined the relative contribution of individual taxa to the total, community-wide trait level from the product of taxon relative abundances and their respective trait abundance distributions (equation 3.3). We refer to the distribution of the relative contribution of each taxon to total, community-wide trait level as the trait contribution distribution (equation 3.4). Thus, an abundant taxon that expresses a small amount of trait per cell might contribute equally to the total trait level within a community as a rarer taxon that expresses a large amount of trait per cell (Figure 3.1B). In cases of presence/absence trait data, the trait contribution distribution reflects the relative abundance of trait-containing taxa.

**Defining and Quantifying Functional Redundancy**

We define functional redundancy as how evenly taxa in a community contribute to the total quantity of trait exhibited by the community. A community in which each taxon contributes an equal quantity of trait would be maximally functionally redundant with respect to the trait. For example, a community of heterotrophs in which each taxon consumes oxygen at an equal rate would be maximally redundant with respect to oxygen consumption. Conversely, if that

community contained a single taxon capable of fixing nitrogen, that community would be minimally redundant with respect to nitrogen fixation.

The trait contribution distribution (equation 3.4) provides an invaluable tool for visualizing trait functional redundancy within a community, as these distributions allow for pairwise comparisons in relative trait contribution by different taxa. Comparative analyses of trait contribution distribution from multiple communities may be difficult due to differences in structure and richness. Similar challenges exist when comparing diversity between taxa abundance distributions (113). Consequently, diversity indices have become elegant solutions for comparing community diversity among different communities, as variations in taxa abundance are collapsed into a singular value. The applicability of diversity indices extends beyond taxa abundance distributions and includes any categorical abundance vector (113). The broad application of diversity indices, along with the fact that diversity indices are used for measuring community evenness (69), prompted us to explore using diversity metrics to quantify functional redundancy.

We quantified the diversity of trait contribution distributions (equation 3.4) using traditional diversity indices (equation 3.5). Numerous diversity indices have been proposed in the literature. Three common measures of diversity, taxa richness, Shannon's diversity index, and Simpson's diversity index, are manifestations of the parametric diversity index, in which the $\alpha$ parameter determines the relative weight placed on rare versus common taxa. Equation 3.5 with $\alpha$=0 becomes the sum of taxa present in a sample, or taxon richness. When $\alpha$=1 or 2, equation 3.5 becomes the equation for Shannon's and Simpson's diversity indices respectively. We adopted the parametric diversity index for this flexibility in weighting the importance of taxa. An

evaluation of their viability demonstrated thermophiles did not contribute to biogeochemical processes (117). Next, we quantified a theoretical maximum diversity for the same trait contribution distributions assuming relative contribution of traits were evenly spread among all taxa (equation 3.6). Community evenness can be derived by taking the ratio of community diversity to the theoretical maximum diversity (69). As such, we normalized diversity of the trait contribution distribution by the theoretical maximum diversity (equation 3.7). We refer to this statistical measure (equations 3.5-3.7) of the trait contribution distributions as trait contribution evenness (TCE). TCE measured of how evenly taxa contribute to a specific trait within the community, or the functional redundancy of the trait. TCE differs from other proposed methods measuring functional diversity, richness, and redundancy as these methods use distance-based methods (105, 107, 108), rely on empirical changes in realized traits with respect to community diversity (101, 111), or pairwise comparisons in trait levels (112, 116) when analyzing the trait contribution distributions. In order to facilitate the use of TCE functional redundancy measurements, we have provided an R script to calculate TCE functional redundancy at https://github.com/taylorroyalty/functional_redundancy_script.

**Influence on Functional Redundancy from Taxa Lacking the Trait**

We quantified the diversity of trait contribution distributions (equation 3.4) using traditional diversity indices (equation 3.5). Numerous diversity indices have been proposed in the literature. Three common measures of diversity, taxa richness, Shannon's diversity index, and Simpson's diversity index, are manifestations of the parametric diversity index, in which the $\alpha$ parameter determines the relative weight placed on rare versus common taxa. Equation 3.5 with $\alpha$=0 becomes the sum of taxa present in a sample, or taxon richness. When $\alpha$=1 or 2, equation

3.5 becomes the equation for Shannon's and Simpson's diversity indices respectively. We adopted the parametric diversity index for this flexibility in weighting the importance of taxa. An evaluation of their viability demonstrated thermophiles did not contribute to biogeochemical processes (117). Next, we quantified a theoretical maximum diversity for the same trait contribution distributions assuming relative contribution of traits were evenly spread among all taxa (equation 3.6). Community evenness can be derived by taking the ratio of community diversity to the theoretical maximum diversity (69). As such, we normalized diversity of the trait contribution distribution by the theoretical maximum diversity (equation 3.7). We refer to this statistical measure (equations 3.5-3.7) of the trait contribution distributions as trait contribution evenness (TCE). TCE measured of how evenly taxa contribute to a specific trait within the community, or the functional redundancy of the trait. TCE differs from other proposed methods measuring functional diversity, richness, and redundancy as these methods use distance-based methods (105, 107, 108), rely on empirical changes in realized traits with respect to community diversity (101, 111), or pairwise comparisons in trait levels (112, 116) when analyzing the trait contribution distributions.

We created a simulation which we increased taxa richness ($n$) while fixing the richness of taxa containing a trait ($\nu$) and the relative trait contribution by these taxa ($\tau$) (equation 3.3-3.4). The parametric diversity index requires selecting a value for the parameter, $\alpha$, which varies the sensitivity that the measured functional redundancy has to community taxa contributing less to the overall community trait level. We performed the simulations such that the parameter, $\alpha$, varied from 0 to 2 for each combination of total taxa richness and trait-containing taxa richness (Figure 3.2). Functional redundancy was less sensitive to increasing taxa richness for larger $\alpha$

values. When the community structures are fixed (equations 3.1 and 3.2, respectively), the functional redundancy values derived from different $\alpha$ values are correlated very closely to one another. To emphasize this point, we plotted select $\alpha$ from contours in Figure 3.2A across all taxa richness (Figure 3.2B-C); these plots were nearly linear. Thus, $\alpha$ represents a parameter which simply augments values of functional redundancy and is not a parameter reflecting ecological significance. We recommend that if influences from high beta-diversity in non-trait containing taxa is not desirable in calculating functional redundancy, then $\alpha$ values closer to 2 should be used in equation 3.5.

When evaluating functional redundancy in real biological systems, it may be desirable to relate functional redundancy to environmental factors such as nutrients or temperature. The sensitivity that functional redundancy relates to these parameters depends on the $\alpha$ parameter (Figure 3.2). Another consideration (beyond functional form) includes the influence that the $\alpha$ parameter has on the range in precision, or dynamic range, of functional redundancy. Specifically, the maximum and minimum values of functional redundancy are one and zero, respectively. However, an increase in precision when converging towards zero may be useful to prevent a loss in functional redundancy sensitivity to exogeneous factors. We evaluated the dynamic range in functional redundancy for 10,000 randomly generated artificial communities as a function of the $\alpha$ parameter. Functional redundancy spanned up to ~10 orders of magnitude depending on the choice in $\alpha$ parameter as well as the fraction of taxa with the trait (Figure 3.3). The functional form for dynamic range varied depending on the fraction of taxa with the trait. Communities where all taxa had the trait behaved had a local maximum while communities with <100% of taxa were asymptotic. Nonetheless, the maximum dynamic range for all considered

scenarios occurred at $\alpha<1$. This implies that a greater range of functional redundancy values can be measured, if the diversity index provides low to moderate weighting for taxa with lower trait contribution towards the overall community trait level. A selection of $1 \geq \alpha \geq 0.5$ appeared adequate if making comparisons between communities with drastically different ratios of trait-containing richness taxa richness ($v$) and taxa richness ($n$).

**Functional Redundancy and Trait Resilience to Taxa Extinction**

One tangible ecological consequence of functional redundancy is increased resilience of biogeochemical functions to taxon extinction. When functions are more broadly distributed across taxa, a function is less likely to disappear due to the extinction of individual taxa. This interpretation is often highlighted in macro-ecological studies (105, 118, 119). Therefore, any measure of functional redundancy should reflect that the probability of a trait being extinguished from a community decreases as the redundancy of that trait within the community increases.

The conceptual relationship between functional redundancy and vulnerability of ecosystem function provides an opportunity to demonstrate the efficacy in TCE. To do this, functional redundancy was calculated for randomly generated communities, which followed lognormal distributions (87, 114) (equation 3.8). Individual taxa were randomly assigned a trait level. We simulated random taxon extinction events in the artificial communities until >95% of the original community total trait was gone. The choice in threshold served the purpose of dealing with communities where all taxa possessed a trait. In such communities, a 100% extinction rate was necessary to properly remove the trait from the community. In real biological systems, an ecosystem function may be disrupted well before all taxa contributing to a trait are extinct. For instance, porewater redox conditions become temporally unstable with minor losses

in organoheterotroph composition (101). For our calculation of equation 3.5, we selected $\alpha$=0.5 because it maximized the observable dynamic range of the functional redundancy parameter (Figure 3.3). We observed a positive relationship between the resilience of a trait to random taxon extinction events and the calculated functional redundancy (Figure 3.4A). In addition to TCE, we evaluated the formulation for functional redundancy presented in (105) to draw comparison. As a general observation, this method performed poorly as no clear functional relationship existed between estimates in functional redundancy and trait resilience to extinction (Figure 3.4B). We suspect the discrepancy in performance between TCE and the method from (105) is explained by TCE's use of diversity to quantify functional redundancy. Specifically, the parametric diversity index (equation 3.5) is sensitive to the shape of the distributions of species abundance and trait contribution. These properties likely manifest in our extinction simulation, which is a hypergeometric probability distribution problem, weighted proportionally to the trait contribution distribution. In contrast, the formulation in (105) is a measure of mean dissimilarity in trait contribution, and does not account for non-normal abundance distribution shapes.

We ensured the relationship between TCE and trait resilience was not an artifact of $\alpha$ selection. This was done by testing a range of $\alpha$ values against trait resilience to taxa extinction (Figure 3.5). The relationship between functional redundancy and resilience to extinction remained positive for all evaluated values of $\alpha$ (0.1, 1, and 2). The functional form (e.g., exponential-like versus linear) between functional redundancy and resilience to extinction as well as variance in residuals differed depending on $\alpha$ value selection. Depending on the work, the relationship between functional redundancy and resilience to extinction may matter, and so flexibility in $\alpha$ selection may prove useful.

**Case Study: Redundancy of Nitrogen Cycle Marker Genes in Marine Microbiomes**

To demonstrate a use of TCE, we measured functional redundancy in the marine nitrogen cycle using genomic data from the *TARA Oceans* project (19). Recently, (65) published 2,361 draft metagenome-assembled genomes (MAGs) which represent genomic content from seven oceanographic provinces, three different depths, and four filter size fractions. Relative abundance data used here were reported in (65) and were determined by the proportion of sequence reads mapping to assembled MAGs. For simplicity, we limited our analysis to only MAGs identified in the bacterial size fraction (<3$\mu$m), which corresponds to free-living microbes. The nitrogen cycle was simplified to eight pathways (115) and included $N_2$ fixation, nitrification, $NO_3^-$ reduction, dissimilatory $NO_3^-$ reduction, assimilatory $NO_3^-$/$NO_2^-$ reduction, $NH_4^+$ assimilation, N remineralization, and denitrification. A general summary of the marker genes (downloaded from Swiss-Prot; see Methods) used for quantifying nitrogen cycle pathways in MAGs is summarized in Table 3.1.

Overall, the functional redundancy for individual traits was consistent with our conceptual understanding of primary nitrogen-transforming pathways in the ocean (Figure 3.6). A clear example was the low functional redundancy in the nitrification pathway. We expected this observation as the primary oceanic nitrifying phylum is *Crenarchaeota* (120). Similarly, the $N_2$ fixation pathway, which is predominantly facilitated by a small group of *Cyanobacteria* (115), exhibits low functional redundancy. Conversely, the uptake of inorganic N via $NH_4^+$ and $NO_3^-$/$NO_2^-$ assimilation are broadly expressed traits and should exhibit higher levels of functional redundancy (115). Lastly, the low functional redundancy in dissimilatory $NO_3^-$ reduction and

$NO_3^-$ reduction were not surprising considering these pathways occur in anaerobic and low-oxygen environments, respectively (115).

Some variance existed across oceanographic provinces and depth for intra-pathway functional redundancy ($\alpha$=0.5). However, inter-pathway variance in functional redundancy was generally larger compared to intra-pathway variance in functional redundancy (Figure 3.6). This suggests that the marine nitrogen cycle marker genes considered here are consistently distributed within communities for MAGs retrieved in Earth's ocean. This was consistent with the lack of relationship between the metagenome companion metadata (e.g., $O_2$, inorganic N, $PO_4^{3-}$, latitude, longitude, etc.) and functional redundancy. This is not particularly surprising, as saline water biomes exhibited the lowest alpha-diversity (not to be confused with the $\alpha$ parameter in equation 3.5) and largely constrained beta-diversity among free-living microbiomes (121). An earlier analysis qualitatively demonstrated similar KEGG ontology in *TARA Oceans* metagenomes, regardless of oceanographic province or depths (19).

In our analysis of *TARA Oceans* metagenomes, traits were loosely defined, such that redundancy in function may be observed based on genomic potential. As a point of emphasis, measurable properties (e.g., nitrogen uptake) provide direct measures of ecosystem function that are more easily integrated into predictive numerical models. In an ideal scenario, the latter would be discernible in all ecological studies, for all ecosystem functions. However, the disproportionate number of lineages that lack cultured representatives highlights the continued importance of nucleotide sequencing in microbial ecology (15, 122). Thus, our observation necessitates the development of ecological theory that is inclusive to potential traits, as well as realized traits. For instance, technological advances with single-cell nanoSIMS may provide a

method for quantifying lineage-specific metabolic rates for *in situ* inoculation studies (123). Coupling these metabolic rates with abundance data, we believe functional redundancy can be measured for realized traits with TCE.

## Methods and Theory

### Definition of Functional Redundancy

Here, we present the underlying theory used to quantify functional redundancy. Functional redundancy was defined as how evenly community members contribute to a net trait within the community. Community members represent the aggregation of individuals within phylogenetic clades of a predefined taxonomic level, and thus, are referred to as taxa from henceforth. As this definition of functional redundancy is a measure of the evenness in trait contribution from individual taxa, we propose adapting traditional approaches of measuring diversity and evenness for quantifying functional redundancy. Specifically, we adopt the parametric diversity index due to the degree of generalizability. Consequently, relative trait contribution by taxa to a net trait on a community level has to be expressed as an abundance distribution (113), or the trait contribution distribution.

Here, we define a microbial community, *C*, with *n* taxa, each with a relative abundance of $p_i$. Thus, the set of taxa in *C* can be described as follows:

$$C = \{p | p_i > 0 \; \forall \; p_i; \sum_{i=1}^{n} p_i = 1\} \quad (3.1$$

where *p* is the set of relative abundances of taxa, such that any given relative abundance for the $i^{th}$ taxon is greater than zero and all taxa relative abundances sum to one. Equation 3.1 represents a traditional abundance distribution (114).

Taxa in *C* cumulatively contribute to a net trait, *T*, which reflects the quantity of a trait on the community level. Units of *T* are flexible such that *T* can be expressed in terms of rates (e.g., g C $m^{-2}$ $year^{-1}$ for net primary productivity), quantities (e.g., gene copies), or dimensionless numbers; however, *T* must be quantitative. Our goal was to express the community, *C* (equation 3.1), in terms of the relative contribution by individual taxa towards a total trait, *T*, at a fixed time and space. This requires knowing the taxa-specific trait level, $t_i$, and the respective taxa relative abundance, $p_i$. If the distribution of taxa-specific trait levels, $t_i$, are known, then relative abundances in equation 3.1 can be substituted with $t_i$ such that:

$$C_t = \{t | t_i \geq 0 \; \forall \; t_i; \sum_{i=1}^{n} t_i = T\} \quad (3.2$$

where $C_t$ is the distribution of *t* across all taxa in a community and the sum of all *t* equals *T*. Thus, the relative contribution that the $i^{th}$ taxon makes towards *T*, defined here as $\tau_i$, was calculated as:

$$\tau_i = \frac{p_i t_i}{\sum_{i=1}^{n} p_i t_i} \quad (3.3$$

As each $\tau_i$ is a fraction of *T*, we reformulate the abundance distribution for the community, *C*, as an abundance distribution with respect to $\tau$ (trait contribution distribution), such that:

$$C_{\tau,all} = \left\{\tau \middle| \tau_i \geq 0 \; \forall \; \tau_i; \sum_{i=1}^{n} \tau_j = 1\right\} \quad (3.4.1$$

and

$$C_\tau = \left\{\tau \in C_{\tau,all} \middle| \tau_i > 0 \; \forall \; \tau_i; \sum_{i=1}^{v} \tau_i = 1\right\} \quad (3.4.2$$

where again, $\tau_i$ is the relative contribution by the $i^{th}$ taxon to *T*, such that all $\tau$ sum to equal 1. The community represented in equation 3.4.1 considers the situation where $\tau$ can equal 0 because the $i^{th}$ taxon has a $p_i$>0 but may not contribute to *T* ($t_i$=0). The community represented in equation

3.4.2 is a subset of equation 3.4.1, where $C_\tau$ contains only taxa from $C_{\tau,all}$ with $t>0$. As such, we adopted the variable, $v$, which represents the richness of taxa in $C_\tau$. Note that equation 3.4.1 and 3.4.2 equal when all taxa contribute to $T$ ($t>0$).

The necessity for differentiating equation 3.4.1 and equation 3.4.2 is that the trait contribution distribution in equation 3.4.2 is consistent with attributes of a community. Specifically, equation 3.4.2 does not include absent taxa which are not incorporated in conventional diversity metrics (113, 114). Thus, we utilized a parametric approach for characterizing parameters of diversity and evenness with respect to equation 3.4.2 (107). Diversity in trait-contributing taxa was determined as:

$$H_\tau = \begin{cases} \frac{1-\sum_{i=1}^{v} {\tau_i}^{\alpha}}{\alpha-1} & \alpha \neq 1 \\ \sum_{i=1}^{v} -\tau_i \log \tau_i & \alpha = 1 \end{cases} \quad (3.5$$

where $\alpha$ is a weighting parameter that defines the importance of taxa contributing less to $T$ compared to taxa which contribute more. When $\alpha$=0, $H_\tau$ simplifies to $v$-1, or the richness of trait-containing taxa minus one. When $\alpha$=1, $H_\tau$ simplifies to Shannon's index. When $\alpha$=2, $H_\tau$ simplifies to Simpson's index. The first scenario represents the least sensitive weighting towards relative trait contribution as it effectively measures how many taxa contain a trait. In contrast, increasing $\alpha$ decreases the sensitivity of $H_\tau$ to taxa which contribute less to $T$. As, our definition of functional redundancy is how evenly different taxa contribute to the community-level total trait within the community, we calculated a theoretical maximum for $H_\tau$ which assumes that trait level, $t_i$, is evenly distributed in equation 3.4 such that:

$$C_{\tau,E} = \left\{ \tau \middle| \tau_i = \frac{1}{n} \; \forall \; \tau_i; \sum_{i=1}^{v} \tau_i = 1 \right\} \quad (3.6$$

Implicitly, equation 3.6 assumes $v=n$, as this is the only scenario where the condition $\sum_{i=1}^{v} \tau_i = 1$ is held true. Thus, $H_E$ is derived by calculating equation 3.5 using equation 3.6. As evenness is the ratio between observed diversity and the theoretical maximum diversity, we calculated functional redundancy, $R$, as:

$$R = \frac{H_\tau}{H_E} \quad (3.7$$

**Functional Redundancy's Dependency on Diversity Continuum**

We wanted to observe the effect that different diversity metrics (i.e., different values of $\alpha$ in equation 3.5) had on $R$. This was performed by calculating $R$ with a fixed community and varying $\alpha$ from 0 to 2. We further characterized sensitivity in $R$ with changes in community richness while fixing the trait-containing taxa (i.e., equation 3.4 constant while equation3.1 changes). This scenario evaluated the sensitivity in $H_E$ to equation 3.5 for different $\alpha$ while maintaining a constant $H_\tau$. $C_\tau$ was defined as a lognormally-distributed community (87) following the functional form:

$$S(r) = S_0 e^{-a^2 r^2} \quad (3.8$$

where $S_0$ was the maximum abundance, $r$ was the $i^{th}$ octave, and $a$ was the inverse width of the lognormal distribution. Values for equation 3.8 were $a$=0.2, $r$ spanned between 0 to 9 ($v$=10), and $S_0$=1. $S(r)$ was then normalized by the sum of abundances across all $r$ so that the condition, $\sum_{i=1}^{v} \tau_i = 1$, was met. Afterwards, we dynamically increased values of $n$ from 10 to 100 (lognormal-evenly distributed increments) for generating the theoretical, maximally diverse trait contribution distributions (equation 3.6). $R$ was then calculated while keeping the $H_\tau$ at a fixed value. This exercise was performed for $\alpha$ spanning from values 0 to 2 at increments of 0.01.

**Functional Redundancy Against Trait Vulnerability Simulations**

We generated artificial communities such that the variables of community richness, diversity (with respect to *p*), and *t* were varied over a dynamic range. Community richness, *n*, was randomly assigned a value from 30 to 100 for individual communities. The relative abundance, *p*, for individual taxa within a community were then defined using equation 3.8. Values of *a* where randomly selected from 0 (even) to 0.4 (operationally more uneven) at 0.01 intervals. After establishing taxa relative abundances, we prescribed levels of trait to individual taxa. Level of trait, *t*, was assigned a value between 0 (does not contain) and 1 (maximum level) at intervals of 0.01. We randomly selected one of three scenarios in the assignment of *t*. The first scenario considered every taxon containing the trait, where every taxon had the trait greater than 0. The second scenario had 50% of the taxa containing the trait at a random level greater than 0 while the other 50% did not contain the trait. The last scenario had 20% of taxa containing the trait at a random level greater than 0 while 80% of taxa did not have the trait. In total, 10,000 artificial communities were generated with this approach. For each artificial community, functional redundancy with equations 3.1-3.8. For comparison, we calculated *R* based on (105) (using code published in Appendix S1). In our calculations of Rao's functional diversity, we utilized Euclidean distances.

We performed two separate simulations with the artificially generated communities described above. The first simulation varied $\alpha$ from 0 to 2 in equation 3.5 to determine which $\alpha$ maximized the dynamic range in observed functional redundancy. In this analysis, we excluded scenarios where $R$=0 and $R$=1. As functional redundancy spanned ~10 orders of magnitude, we reported the range after taking the $\log_{10}$ functional redundancy. The second simulation caused

random taxon extinction events within the artificial communities described above. $C_\tau$ was determined for individual artificial communities with equations 3.1-3.4. Taxon extinction events were simulated by randomly removing the same taxa from $C$ (equation 1) and $C_\tau$ (equation 3.4). If a taxon in $C$ did not contain the trait, then no taxa were removed from $C_\tau$. After each taxon extinction event, the total $\tau$ in $C_\tau$ was determined by summing the remaining elements within $C_\tau$. Random taxon extinction events continued until the sum of $\tau$ remaining in $C_\tau$ was >95% from starting value of $\tau$ (i.e, $\tau$=1). Trait robustness to extinctions was then defined as the fraction of taxon richness ($n$) which went extinct prior to 95% of the trait being lost. The extinction simulation was performed 100 times for each of the individual 10,000 artificial communities. The trait robustness generated from the 100 simulations were then averaged for each of the individual 10,000 artificial communities.

**A Case Study of Functional Redundancy in the Global Ocean Nitrogen Cycle**

Draft metagenome-assembled genomes (MAGs) from (65) were downloaded from NCBI (38). In addition, MAG completeness, MAG relative abundance estimates, ocean depths, and ocean regions, defined in (65), were directly downloaded from Figshare (https://doi.org/10.6084/m9.figshare.5188273). We tabulated nitrogen-cycle marker genes in individual MAGs by blasting MAG proteins against nitrogen-cycle marker gene databases using the BLASTP algorithm ($E$-value=$10^{-10}$) (22). Nitrogen cycle marker genes were categorized into eight categories: $N_2$ fixation (*nifH*), nitrification (*amoA*, *hao*), $NO_3$ reduction (*napA*, *narG*), dissimilatory $NO_3$ reduction (*nrfA*), assimilatory $NO_3$/$NO_2$ reduction (*narB*, *nasA*, *nasB*, and *nirB*), $NH_4^+$ assimilation (*gdhA*, *glnA*), N remineralization (*gltB*), and denitrification (*nirK*, *nirS*, *norB*, and *nosZ*) (115). Individual BLAST databases were constructed for each nitrogen cycle

category. Gene names associated with individual categories were searched for in the Swiss-Prot database (124). All proteins matching gene names were subsequently downloaded and used in BLAST database construction. The number of matching genes for individual MAGs were normalized by the MAG completeness (fraction) to account for incomplete binning and assembly. Functional redundancy was then calculated for individual stations, at individual depths, for each nitrogen cycle category using equations 3.1-3.7.

# Appendix Three

## Tables

**Table 3.1.** A summary of marker gene sequences used in the BLASTP analysis of *TARA Oceans* MAGs. All marker genes were downloaded from Swiss-Prot.

| Pathway | Gene Names | EC Number | # Marker Genes |
|---|---|---|---|
| N Remineralization | *gltB* | 1.4.1.13 | 14 |
| N Fixation | *nifH* | 1.18.6.1 | 127 |
| Nitrification | *amoA* | 1.14.99.39 | 2 |
| | *hoa* | 1.7.2.6 | 1 |
| $NO_3^-$ Reduction | *napA* | 1.9.6.1 | 136 |
| | *narG* | 1.7.5.1 | 13 |
| Dissimilatory $NO_3^-$ Reduction | *nrfA* | 1.7.2.2 | 77 |
| Assimilatory $NO_3^-$/$NO_2^-$ Reduction | *narB* | 1.7.5.1 | 3 |
| | *nasA* | 1.7.7.2 | 1 |
| | *nasB* | 1.7.1.4 | 3 |
| | *nirB* | 1.7.1.15 | 5 |
| $NH_4^+$ Assimilation | *gdhA* | 1.4.1.3 | 14 |
| | *glnA* | 6.3.1.2 | 173 |
| Denitrification | *nirK* | 1.7.2.1 | 7 |
| | *nirS* | 1.7.2.1; 1.7.99.1 | 4 |
| | *norB* | 1.7.2.5 | 2 |
| | *nosZ* | 1.7.2.4 | 11 |

## Figures

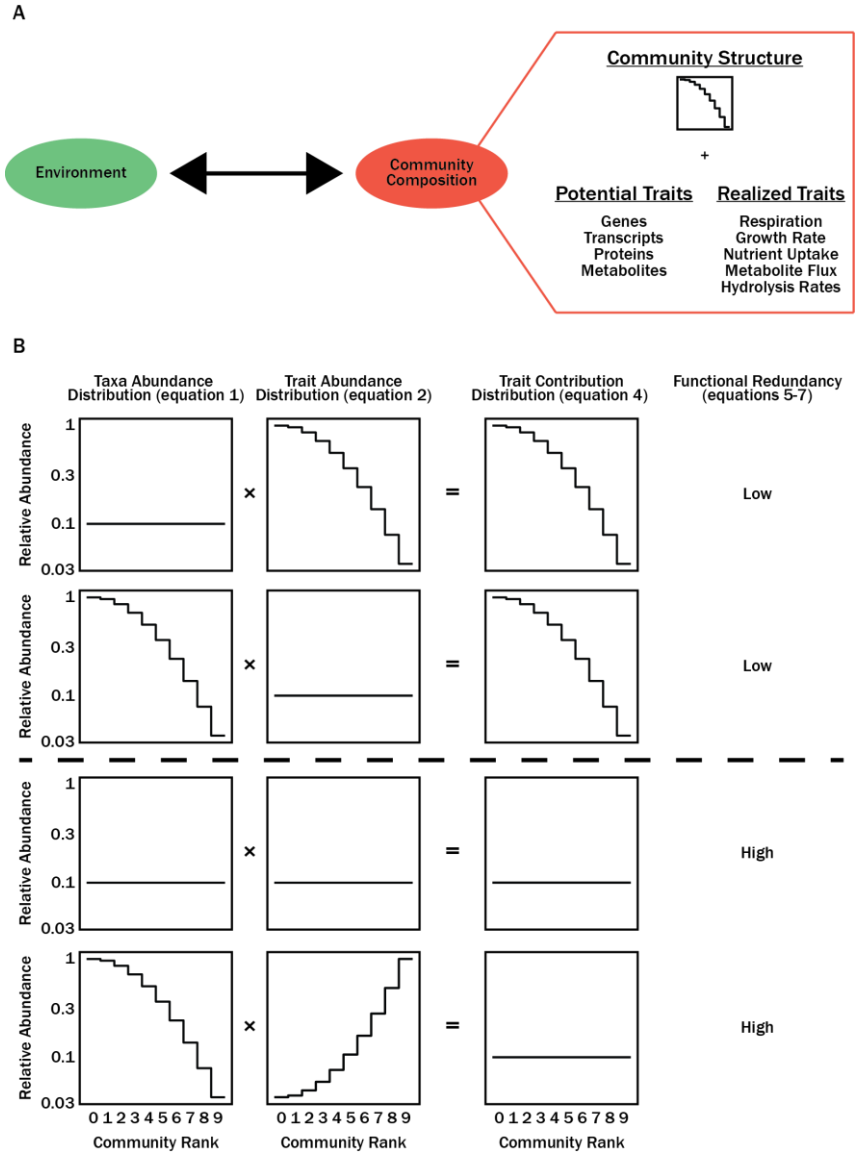


**Figure 3.1.** The interplay between environmental parameters and microbial community composition requires an understanding of both community taxa abundances as well as trait levels for respective taxa (A). Community taxa abundance and relative trait level are scalable and provide a measure for the contribution that different taxa have to the total community level trait (B).

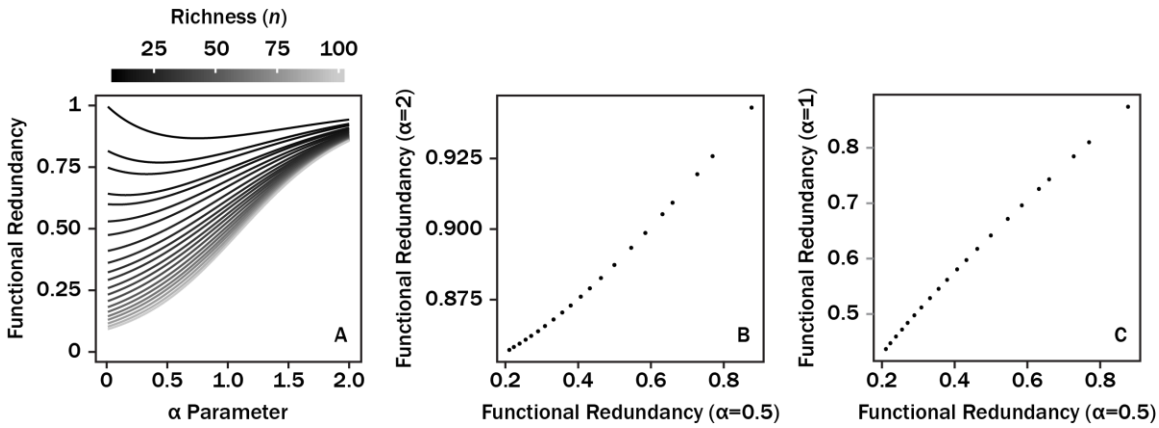


**Figure 3.2.** The sensitivity in functional redundancy in response changing $\alpha$ in equation 3.5 (A). Contours correspond to increasing community richness ($n$) while maintaining the structure of the trait-containing community. Comparisons between contours from (A) when $\alpha$=0.5 and $\alpha$=2 (B) as well as when $\alpha$=0.5 and $\alpha$=1 (C).

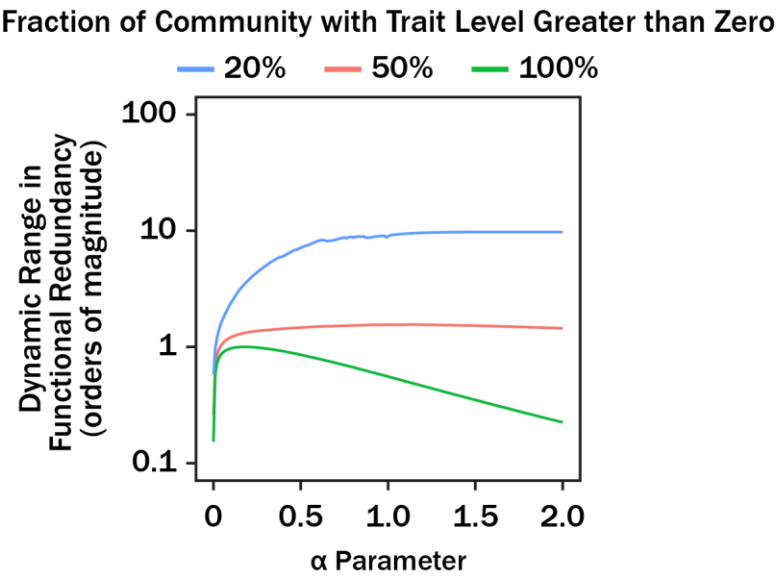


**Figure 3.3.** The dynamic range in functional redundancy for randomly generated artificial communities as a function of $\alpha$ selected for equation 3.5. Contours correspond to the fraction of the community containing a trait. Note that at $\alpha = 1$ and the fraction of community was 20%, the dynamic range was smoothed due to precision error from the log-transformation in equation 3.5 when calculating function redundancy. All other data reflect results from artificial communities.

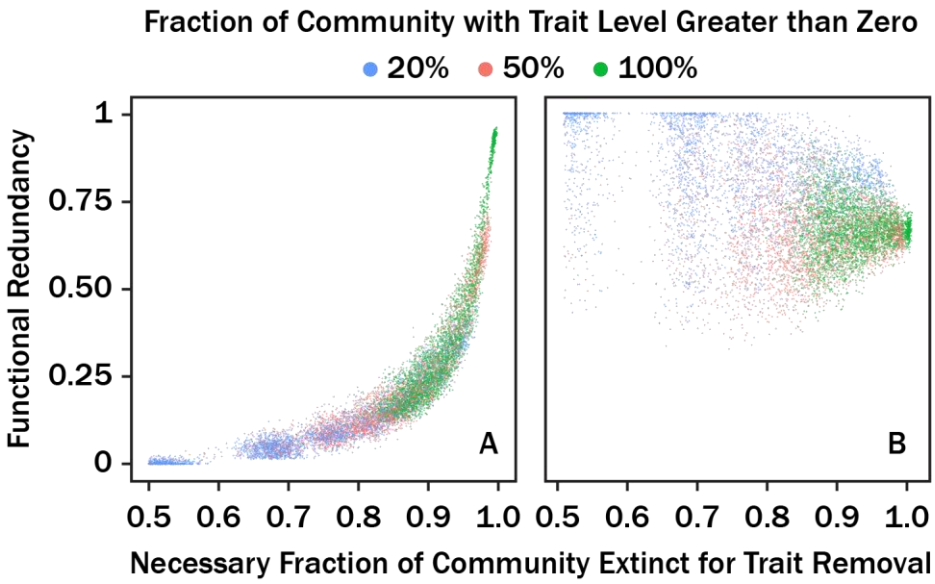


**Figure 3.4.** A comparison of functional redundancy versus the fraction of community richness that goes randomly extinct prior to a 95% of trait level being lost on the community level. Functional redundancy was calculated with equations 3.1-3.7 ($\alpha$=0.5) (A) as well as using the method in (105) (B).

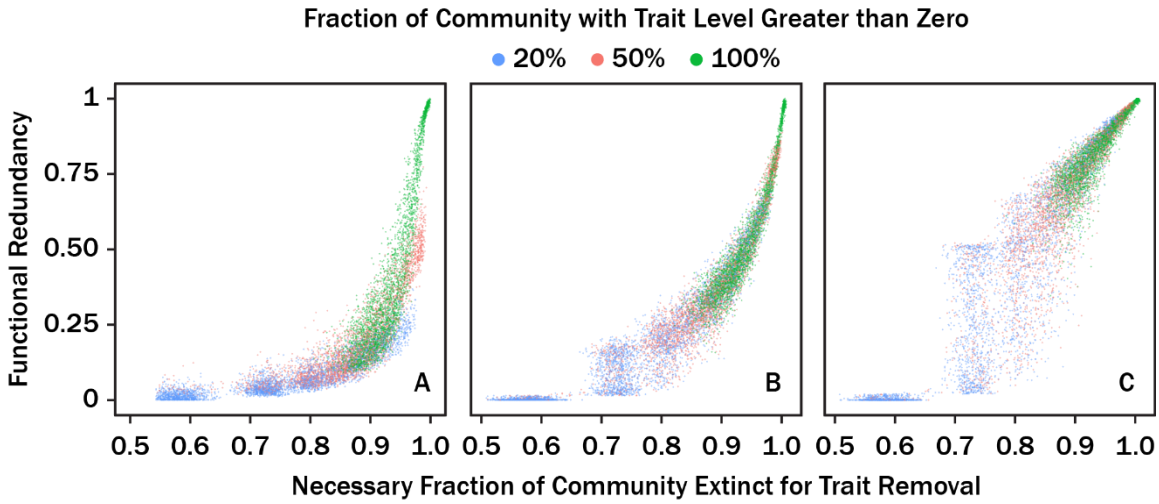


**Figure 3.5.** Same as Figure 3.4A but functional redundancy was calculated with equations 3.1-3.7 such that α=0.1 (A), α=1 (B), and α=2 (C).

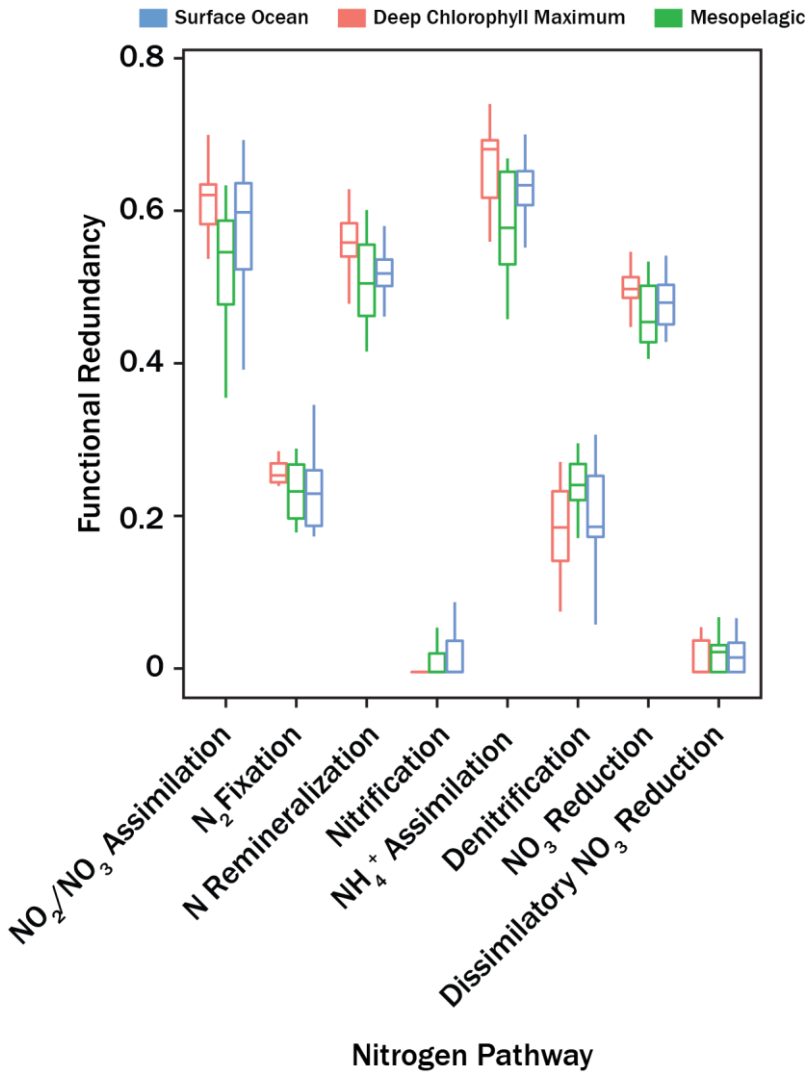


**Figure 3.6.** Boxplots showing functional redundancy calculated for eight nitrogen-transforming pathways in *TARA Oceans* draft MAGs. Different colors correspond to sample depth. Whiskers correspond to 150% of the inter-quantile region.

# CHAPTER FOUR

# SULFUR AND CARBON BUDGETS IN THE WHITE OAK RIVER ESTUARY BASED ON REACTION-TRANSPORT MODELING

## Abstract

Estuary sediments at White Oak River (WOR), North Carolina (USA), experience high sedimentation rates, with a high organic carbon (OC) fraction, and harbor diverse microbial metabolisms. Rapid zonation in reduction-oxidation conditions suggest the WOR may be a suitable model system for microbial niche modeling; however, additional geochemical and hydrological characterization is necessary first. Here, I report measurements of major seawater ion concentrations in porewater, ammonium porewater concentrations, phosphate porewater concentrations, total organic carbon (TOC), total organic nitrogen (TON), bulk $\delta^{13}C$, porewater methane concentrations ($CH_4$), and dissolved organic carbon spectral properties. The select geochemical parameters were related to the microbial metabolic transformations and used for recalculating budgets for dissolved inorganic carbon (DIC), TOC, $CH_4$, and sulfate ($SO_4^{2-}$). Consistent with earlier observations, I found evidence for aerobic, $SO_4^{2-}$ reducing, fermentative, anaerobic oxidation of methane (AOM), and methanogenesis metabolisms in Station H sediments, a site extensively characterized in the literature with respect to WOR estuary microbiology. Notably, TOC concentrations were elevated at the sulfate-methane transition zone (SMTZ). I hypothesized that increased TOC concentration was from biomass production in microbes utilizing the AOM metabolism as bulk $\delta^{13}C$ and C:N ratios of organic fractions concurrently decreased and increased, respectively. Lastly, budgets were comparable with previous estimates when assuming diffusion-dominated chemical transport. The TOC burial, DIC respiration, $CH_4$ production, and $SO_4^{2-}$ reduction fluxes were 6091, 3178, 35, and 1977 mmol $m^{-2}$ $yr^{-1}$, respectively. Unlike previous budgets analyses, I considered advection-dominated chemical transport (groundwater upwelling <20 cm $yr^{-1}$). I found $CH_4$ oxidation and $SO_4^{2-}$

reduction rates increased and decreased, respectively, with increasing groundwater upwelling. When groundwater upwelling rates were 14.9 cm $yr^{-1}$, net $SO_4^{2-}$ reduction fluxes were predicted to be 0 mmol $m^{-2}$ $yr^{-1}$. Groundwater upwelling >0 cm $yr^{-1}$ suggested cryptic sulfur cycling is necessary to maintain steady-state $SO_4^{2-}$ concentration profiles. Although carbon and sulfur budgets appear consistent on decadal timescales, the SMTZ depth varies among earlier studies. Variables influencing variability in geochemistry on short and intermediate timescales need constraint to better contextualize microbial ecological findings with respect to time.

## Importance

High nutrient estuary sediments support a diverse suite of microbial metabolisms. The combination of diverse microbial metabolisms and ease of access may make high nutrient estuary sediments good model systems for studying niche partitioning theory. Nonetheless, characterizing the temporal stability in geochemistry is an important prerequisite for properly contextualizing ecological findings. Here, I reevaluated elemental budgets for carbon and sulfur at the White Oak River and compared our results to budgets reported >20 years ago. Overall, my findings suggest that the net carbon and sulfur budgets are comparable to one another on decadal timescales; however, a literature survey of studies measuring sulfate and methane profiles at the White Oak River suggest individual geochemical parameters vary on short and intermediate timescales. Thus, future ecological studies should properly consider short and intermediate timescale processes influencing both ecology and geochemistry.

# Introduction

Conventional wisdom dictates that electron acceptors are sequentially depleted in marine sediments based on energy yield from microbially-mediated reduction-oxidation (redox) reactions. Assuming the dominant electron donor is organic matter (OM), the sequence that electron acceptors are depleted is $O_2$, $NO_3^-$, Mn(IV), Fe(III), $SO_4^{2-}$, and $CO_2$ (125). Sediment environments with high depositional OM fluxes lead to rapid depletion of electron acceptors and rapid transition between dominant redox metabolisms (126). Thus, studies requiring a large range of microbial metabolic diversity could leverage sediments with high OM depositional fluxes as model study systems.

One sediment system fitting this criteria includes blackwater river estuaries on the East coast of the United States (127). Termed "blackwater river" due to high humic concentrations that result in black colored water (128), blackwater estuaries have high terrestrial OM runoff, rapid autochthonous OM production, and $SO_4^{2-}$ from seawater intrusion, estuary systems along the Coastal Plains support a wide range of microbial metabolisms (17). In particular, the White Oak River (WOR) estuary has received considerable research interest related to sediment geochemistry and microbiology during the last half century (17, 34, 35, 129–134). Porewater geochemical analyses indicate that all major electron acceptors ($O_2$, $NO_3^-$, Mn(IV), Fe(III), $SO_4^{2-}$, and $CO_2$) are observable at <100 cmbsf (129). For these reasons, I selected the WOR as a study site for modeling microbial trait distributions with respect to geochemical gradients (135, 136).

Although our primary goal was modeling microbial trait distributions, a budget analysis of carbon and sulfur in WOR sediments had not been evaluated since 1998, or following the accumulation of 6.8 cm sediment—assuming a 0.31 cm $yr^{-1}$ rate for the WOR estuary (34, 35,

130). Specific interest in carbon and sulfur pertains to these elements being the primary source of electron donors and acceptors in WOR sediments (130). In these studies, $CH_4$, $CO_2$, and $SO_4^{2-}$ concentration profiles were assumed to be in steady-state, with diffusion-dominated transport. Here, major seawater ion concentrations, total organic carbon (TOC), total organic nitrogen (TON), dissolved organic matter (DOM) absorption spectra, $NH_4^+$ concentrations, $PO_4^{3-}$ concentrations, $CH_4$ concentrations, and bulk organic $\delta^{13}C$ values were measured in a 62 cm sediment core collected on 29 May 2019. I constructed budgets for TOC, dissolved inorganic carbon (DIC), $CH_4$, and $SO_4^{2-}$ using a reaction-transport model that assumed steady-state concentrations (35). When assuming chemical transport was diffusion-dominated, the annual fluxes for TOC, DIC, $CH_4$, and $SO_4^{2-}$ were comparable to previously reported values (35). However, I noted a few differences in the concentration depth profiles among earlier studies. Notably, there was a previously unreported increase in TOC at the sulfate-methane transition zone (SMTZ) which may reflect biomass growth from an anerobic oxidation of methane (AOM) metabolism (137). Furthermore, advection-dominated chemical transport resulted in elevated $CH_4$ oxidation rates and reduced $SO_4^{2-}$ reduction rates. Based on the findings here, assumptions of steady-state may be valid for WOR sediments; however, better constraint on groundwater upwelling is necessary to better characterize sulfur cycling in WOR sediments.

## Materials and Methods

### Site Description

The White Oak River (WOR) is an ~80 km long blackwater river located in eastern North Carolina (USA) (Figure 4.1). The WOR is a part of a broader network of incised river valleys in the United States Coastal Plains, which are characterized by a high sediment OC fraction and

high sedimentation rates (127). Water column temperatures vary from 22°C to 32°C during summer months and have been observed to be as low as 3.5°C during winter months (130, 138). The WOR is summarized as three major sections: 1) the upper estuary that is characterized by a narrow, meandering freshwater river; 2) the middle estuary that is characterized by the widening of the river and high variability in water column salinity controlled by tidal patterns; and 3) the lower estuary that is characterized by a deltaic section with high water column salinity (133). Sediment OC fraction decreases while sediment porewater $SO_4^{2-}$ concentrations increase as the WOR transitions from upper to lower estuary. The nature of these gradients results in methanogenesis being the dominant microbial metabolism in the WOR upper estuary sediments and $SO_4^{2-}$ reduction being the dominant microbial metabolism in the WOR lower estuary sediments (130).

The hydrological flow in WOR sediments is influenced by groundwater discharge from the underlying Castle Hayne Aquifer. Groundwater discharge estimates throughout the WOR estuary ($Cl^-$ concentration gradients) indicate heterogenous discharge spanning 0.6 to 40 cm $yr^{-1}$ (133). A geological survey of aquifers across the North Carolina Coastal Plains indicates the confining unit for the Castle Hayne Aquifer is often thin (4 feet thick), and in some cases, completely absent. The latter scenario leads to direct contact between the Castle Hayne Aquifer and river systems (139). To my knowledge, direct estimates of sediment column thickness are lacking in the literature for the WOR estuary. However, acoustic seismic measurements indicate sediment thickness is approximately 3.5±0.5 m in the lower estuary while upper and middle estuary sediment thickness was not resolvable due to seismic wipeouts from interstitial $CH_4$ (133).

## Core Sampling

A small boat was sailed on the WOR Station H (34° 44.490'N, 77° 07.44'W) (Figure 4.1.) on 29 May 2019. The selection of Station H was based on extensive microbiological characterization performed in earlier studies (e.g., 8, 117–119). Multiple sediment cores were collected using a push core technique. The sediment core analyzed here was ~62 cm long. The sediment core was transported back to the University of Chapel Hill's Institute of Marine Sciences (IMS), Morehead City, NC, at ambient temperature. A schematic showing sample preparation for these measurements is provided in Figure 4.2. Upon arriving at IMS, the push core was immediately sectioned at 1 cm intervals in a wet chemistry lab. For individual depth intervals, sediment was sub-cored and partitioned into two sample aliquots. The first aliquot was 5 mL and was used for OM and porewater ion analysis. The second aliquot was 3 mL and used for $CH_4$ concentration analysis. The $CH_4$ aliquots were stored in crimped-sealed serum vials. The serum vials contained 1 mL 0.1 M KOH. The purpose of the KOH was to kill all living cells. After the sediment was added to the serum vial, the $CH_4$ aliquot was shaken for ~1 min to ensure KOH homogeneity. The OM/IC aliquot was stored in 15 mL falcon tubes. After samples were allocated into respective storage containers, all samples were stored in coolers with dry ice. Samples were transported back to the University of Tennessee, Knoxville (UTK) on 30 May 2019, and stored in -80°C freezers until sample analysis.

## Methane Analysis

The sediment aliquots designated for $CH_4$ were unfrozen and 0.5 mL of headspace gas was sampled from individual serum vials. The sample headspace gas for an individual sample was injected into a gas chromatograph (Agilent 7890 GC). $CH_4$ was detected using a flame

ionization detector**.** $CH_4$ concentrations (ppmv) were quantified from a five-point calibration curve (created 16 May 2016) that was generated from known $CH_4$ concentrations ($R^2$=0.99). $CH_4$ concentrations were converted into units of molarity as follow (131):

$$[CH_4] = \frac{p(CH_4)V_{headspace}}{RT\varphi V_{sediment}1000(1-\beta)} \quad (4.1$$

where [$CH_4$] was concentration (mM; with volume in respect to wet sediment), $p(CH_4)$ was the measured partial pressure of $CH_4$ (ppmv), $V_{headspace}$ was the serium vial headspace volume (mL) after the addition of sediment and KOH, $R$ was the universal gas constant (L atm $mol^{-1}$ $K^{-1}$), $T$ was the temperature (K) that $CH_4$ was allowed to reach equilibrium between the serum vial headspace and sediment, $\varphi$ was sediment porosity, $\beta$ is the Bunsen solubility coefficient, and $V_{sediment}$ was the volume (mL) of sediment in the serum vial. $CH_4$ reached equilibrium at $T$=294 K. Sediment porosity in WOR sediments are invariant with depth. Thus, a constant value of $\varphi$=0.81 was used for equation 4.1. The Bunsen solubility coefficient represents the solubility of $CH_4$ in water and has been parameterized as a function of salinity and temperature (140). Here, $\beta$=0.002 was used based on the temperature.

**Porewater Chemistry**

Porewater was isolated by centrifuging 5 mL of sediment slurry for individual depth intervals at 4500 x g for 10 minutes. The porewater overlying sediment was extracted using a syringe needle. Water close to the water-air interface was not included in the porewater sample. This minimized the collection of porewater where sulfide was partly oxidized from ambient $O_2$ (visually apparent layers). The extracted porewater was filtered through a 0.22 µm filter. Filtered porewater was split into three aliquots. The first aliquot (1.9 mL) was prepped specifically for $SO_4^{2-}$ analysis with ion chromatography (IC). The second aliquot (0.1 mL diluted in 0.9 mL

milli-Q $H_2O$) was allocated for inorganic cation and anion (excluding $SO_4^{2-}$) analysis IC. The third aliquot (1.5 mL) was for dissolved organic matter (DOM) optical analysis. The remaining sediment pellets were stored for OM analysis (see below).

**Ion Chromatography**

The $SO_4^{2-}$ aliquot (see above) was acidified with 0.1 mL of 10% HCl. The HCl partitioned sulfide species into $H_2S$. The headspace for the porewater was open to ambient conditions for 1 hour to allow outgassing of $H_2S$. The residual sulfur was interpreted as *in situ* $SO_4^{2-}$.

IC measurements were performed at the UTK Stable Isotope Lab service facility (https://eps.utk.edu/research/facilities_isotope.php). Inorganic ion concentrations for $SO_4^{2-}$, $Na^+$, $Cl^-$, $Mg^{2+}$, $K^+$, $PO_4^{3-}$, $Br^-$, $NH_4^+$, and $Ca^{2+}$ were measured on a Dionex ICS 2000 (cations) and a Dionex ICS 2100 (anions). The detection limit for ions was determined by taking a regression from depth intervals 40 to 60 cmbsf. As residuals were normally distributed, the 95$^{th}$ percentile of residual absolute values were taken as the detection limits for each ion species. Prior to sample injection, the anion/cation and $SO_4^{2-}$ aliquots were diluted by a factor of 10 to prevent peak saturation. Both cation/anion and $SO_4^{2-}$ aliquots were corrected for factor of 10 dilution effect in reported concentrations. In addition, the $SO_4^{2-}$ aliquot was corrected for the dilution effect from the 0.1 mL HCl.

Ion concentration profiles exhibited a change in gradient with depth at 15 cmbsf (see Results). The change in sampling angle necessary to explain the change in concentration gradients was determined using a 2 x 2 rotation matrix, such that:

$$\begin{bmatrix} cos(\theta)x_0 - sin(\theta)y_0 \\ sin(\theta)x_0 + cos(\theta)y_0 \end{bmatrix} = \begin{bmatrix} x_{obs} \\ y_{obs} \end{bmatrix} \quad (4.2$$

where $x_0$ and $y_0$ are elements of the sample vector (i.e., orientation of the core with respect to the sediment-water interface), $\theta$ is a bias in sample angle, and $x_{obs}$ and $y_{obs}$ define the observed sample vector given a bias in sample angel θ. Here, $x_0$ and $y_0$ were 0 cm and 1 cm, respectively (assumes sampling perpendicular to sediment-water interface). Values for $\theta$ were varied from 0 to 90° at 0.1° intervals. The calculated $y_0$ values correspond to the biased sample depth with respect to the sediment-water interface. I calculated the average change in concentration over a 1 cm interval by taking regressions from 3 to 15 cmbsf for all seawater ions. These represent the unbiased concentration gradients. I then predicted observed concentrations with the resultant model slopes (3 to 15 cmbsf regression) using the $y_0$ values. Because the predicted concentrations correspond to changes in concentration over 1 cm intervals along the sample core, the predicted concentration (minus the intercept) was interpreted as the biased slope. I did this for all sampling angles, $\theta$. The closest biased slope to the observed slopes from 16 to 60 cmbsf (based on linear regression) were taken as the sampling angles necessary to change the concentration slopes at 15 cmbsf.

**DOM Absorption Spectra**

UV-VIS absorbance spectra were measured for porewater DOM at each depth interval using a UV-VIS spectrophotometer (Ocean Optics FLAME-S-UV-VIS). Ten absorbance spectra were measured and averaged together for each depth. The integration times for the individual spectra were 1000 μs. Absorbance spectra were collected and analyzed at wavelengths from 220 nm to 800 nm. Porewater isolated from sediments (see *Porewater Geochemistry*) from individual depths were filled into a single quartz glass cuvette. After obtaining average absorbance spectra for a given depth interval, the single quartz glass cuvette was washed 15 times with ~1 mL Milli-

Q $H_2O$. After the Milli-Q $H_2O$ rinse, an absorbance spectrum was collected for Milli-Q $H_2O$ in the same fashion as the sediment porewater. The Milli-Q spectra provided reference spectra for absorbance of the sample medium (i.e., $H_2O$).

Absorbance spectra were converted into absorption spectra. Absorbance spectra were first normalized by their respective Milli-Q $H_2O$ spectra, such that:

$$T(\lambda)_i = \frac{I(\lambda)_{sample,i}}{I(\lambda)_{reference,i}} \quad (4.3$$

where $T(\lambda)_i$ is the light transmittance for wavelength, $\lambda$, at the $i^{th}$ depth interval. Transmittance was then used to calculated absorption such that:

$$a(\lambda)_i = \frac{ln(T(\lambda)_i{}^{-1})}{l} \quad (4.4$$

where $a(\lambda)_i$ is the absorption coefficient ($m^{-1}$) for wavelength, $\lambda$, at the $i^{th}$ depth interval, $T(\lambda)_i$ is the same as before, and $l$ is the cuvette pathlength. Here, $l$ was 0.01 m. The resultant absorption spectra were analyzed for the spectral slope ($S_{275\text{-}295}$ nm: $S_{350\text{-}400}$ nm), a proxy used for comparing fulvic to humic acid ratios of DOM end-members (141).

**Total Organic Carbon, Total Organic Nitrogen, and Bulk $\delta^{13}C$**

Measurement and sample preparation for total organic carbon (TOC), total nitrogen (TN), and bulk $\delta^{13}C$ was performed at the UTK Stable Isotope Lab. The sediment pellets from centrifugation (see *Porewater Geochemistry*) were freeze dried for 72 hours (Labconco FreeZone 6 Freeze Dry System) following the manufacture's guidelines. After the sediment dried, ~20 mg of dried sediment was weighed on a Mettler Toledo AX26 analytical balance. Samples were then treated with concentrated HCl to remove inorganic carbon (e.g., carbonate minerals). Single drops of concentrated HCl were added to individual dried sediment samples

with a glass pasteur pipette. After the sediment dried, ~20 mg of dried sediment was weighed on a Mettler Toledo AX26 analytical balance. Samples were then treated with concentrated HCl to remove inorganic carbon. Single drops (~50 uL volume) of concentrated HCl were added to individual dried sediment samples with a glass pasteur pipette. After the addition of a concentrated HCl drop, a sample was visually inspected for evidence of $CO_2$ outgassing (bubbles). If outgassing occurred, an additional drop of concentrated HCl was added to the sediment sample. This process continued until outgassing was not visually apparent in the samples (typically 2-5 drops). The acidified sediment samples were stored at ambient conditions for 24 hours to allow for HCl to react with residual inorganic carbon. Afterwards, samples were transferred to an oven and allowed to dry at 60°C for another 24 hours. Last, the dried sediment samples were analyzed for TOC, TN, and bulk $\delta^{13}$C. Isotopic data were measured on a Costech Elemental Analyzer ECS4010. The $\delta^{13}$C results were reported in units of per mil (‰) in reference to Vienna Pee Dee Belemnite, and assumed to represent a bulk C isotope composition of organic fraction present in the sediment.

TOC and TN measurements were converted from fraction of dry sediment mass to moles $cm^{-3}$ as follow:

$$[TOx] = \frac{(1-\varphi)TOx_{frac}\rho_{dry}}{X_{mw}} \quad (4.5$$

where [TOx] is concentration of either TOC or TN in moles $cm^{-3}$ of wet sediment, $TOx_{frac}$ is the fraction of dry sediment that is either carbon or nitrogen (g $g^{-1}$ dry sediment), $\varphi$ is sediment porosity, $X_{mw}$ is the molecular weight of carbon or nitrogen (12 C g $mol^{-1}$ and 14 N g $mol^{-1}$, respectively), and $\rho_{dry}$ is the density (g $cm^{-3}$) of dry sediment for WOR sediments. Here, I assumed $\rho_{dry}$=2.5 g $cm^{-3}$ (131, 142). No measurable concentrations of $NO_3^-$ and $NO_2^-$ were

present in the sediment core. Thus, TN was converted to TON by assuming all inorganic nitrogen was attributed to $NH_4^+$, and TN and $NH_4^+$ were used to calculate TON by:

$$[TON] = [TN] - [NH_4^+] \quad (4.6$$

**Reaction-Transport Model**

I used a one-dimensional reaction-transport model to infer oxidation and production rates for $CH_4$ and $SO_4^{2-}$ at the WOR (131). The reaction-transport model used conservation equations (143):

$$\frac{\partial C_i \varphi}{\partial t} = D_i(\varphi) \frac{\partial}{\partial z}\left(\frac{\partial C_i \varphi}{\partial z}\right) - \frac{\partial C_i \varphi w}{\partial z} + \sum R_i \quad (4.7$$

where $C_i$ was concentration (moles $cm^{-3}$) for the $i^{th}$ species, $\varphi$ was sediment porosity, $D_i(\varphi)$ was the porosity-corrected diffusion coefficient ($cm^{-2}$ $s^{-1}$) for the $i^{th}$ species, $z$ was depth (cm), $w$ was net vertical advection (cm $s^{-1}$), $\sum R_i$ was the net reaction (moles $cm^{-3}$) for the $i^{th}$ species, $t$ was time (s), and $\partial$ corresponded to the change in the respective variable. Porosity for WOR sediment was found to be invariant with depth, and thus I assumed $\varphi$=0.81 for all depths (131).

Assuming a steady-state chemical profile (i.e., $\frac{\partial C_i \varphi}{\partial t} = 0$), the net rate of reaction, $\sum R_i$, was inferred because both the diffusion and advection terms were known. Noise in the observed chemical data was filtered by fitting a generalized additive models (GAM) to both $CH_4$ and $SO_4^{2-}$ depth profiles (59). The smoothing parameter (sp) was set to 0.1 for both gam fits. The smoothed fits were used in defining $\frac{\partial C_i \varphi}{\partial z}$ for both the diffusion and advection terms. For the diffusion term, the diffusion coefficients for $CH_4$ and $SO_4^{2-}$ were calculated based on experimental data in (144) and from (143), respectively. For these calculations porewater temperature was assumed a

temperature of 23°C (138). The resultant diffusion coefficients were subsequently corrected for porosity (143):

$$D_i(\varphi) = \frac{D_{i,free}}{(1-\log(\varphi^2))} \quad (4.8$$

where $D_i(\varphi)$ was the corrected diffusion coefficient for the $i^{th}$ species, $D_{i,free}$ was the molecular diffusion coefficient for the $i^{th}$ species, and $\varphi$ was porosity. The final diffusion coefficients used in equation 4.7 for $CH_4$ and $SO_4^{2-}$ were $11.2\times10^{-6}$ $cm^{-2}$ $s^{-1}$ and $7.12\times10^{-6}$ $cm^{-2}$ $s^{-1}$, respectively. For the advection term, net vertical advection, $w$, was defined as the summation of sediment burial rate, $\omega$, and groundwater upwelling, $u$, such that:

$$w = \omega + u \quad (4.9$$

The burial rate was assumed a constant value of 0.31 cm $yr^{-1}$ (142). Previous geochemical modeling studies assumed burial rates, $\omega$, dominate vertical advection rates, $w$, at the WOR (34, 35, 130, 131); however, a thermal gradient analysis (138) and $Cl^-$ transport modeling suggest groundwater upwelling may be important in WOR sediments. Geological surveys of North Carolina Coastal Plain aquifers indicates that the confining unit for the Castle Hayne Aquifer, which underlies the WOR estuary, is thin (four feet thick) and sometimes completely missing which leads to direct contact between the aquifer and river systems (139). Thus, I explored the sensitivity of the reaction term in equation 4.7 by varying groundwater upwelling velocities from 0 to 20 cm $yr^{-1}$, where groundwater upwelling is assumed to be sourced from the Castle Hayne Aquifer. These rates are consistent with values previously reported in WOR sediments (133).

We applied the second order central difference formula (143) to solve $\sum R_i$ when assuming steady state concentration (GAM fits) in equation 4.7:

$$\sum R_i = -D_i(\varphi)\frac{\partial}{\partial z}\left(\frac{\partial C_i\varphi}{\partial z}\right) + \frac{\partial wC_i\varphi}{\partial z} \quad (4.10.1$$

and thus,

$$\sum R_{i,z} = -\varphi D_i \frac{C_{i,z+1}-2C_{i,z}+C_{i,z-1}}{z^2} + \varphi(\omega + u)\frac{C_{i,z+1}-C_{i,z-1}}{2z} \quad (4.10.2$$

Lower and upper boundary conditions for dissolved $CH_4$ and $SO_4^{2-}$ were prescribed based on predictions with the GAM fits at 0 cmbsf and 100 cmbsf, respectively. The lower and upper concentration boundary conditions for dissolved $CH_4$ were $1.07\times10^{-2}$ mM and 1.10 mM. The lower and upper concentration boundary conditions for $SO_4^{2-}$ were 0.14 mM and 13.8 mM. Background $SO_4^{2-}$ concentrations for the Castle Hayne Aquifer based on aquifer well water data were comparable to the lower boundary conditions predicted by the fits (see Results). As such, the dependency that $\sum R_i$ had on groundwater upwelling could be determined as all other terms were held constant.

**Total Organic Carbon, Methane, and Sulfur Budgets**

I estimated the sink, or consumption, for TOC, $CH_4$, and $SO_4^{2-}$ in WOR sediments at Station H. The consumption of $CH_4$ and $SO_4^{2-}$ was calculated as:

$$S_i = \Delta z \sum_{z=Z_{i,min}}^{Z_{i,max}} R_i(z) \quad (4.11$$

where $S_i$ was the net sink (nmol $cm^{-2}$ $d^{-1}$) for the $i^{th}$ species, $\Delta z$ was the sample depth interval (cm), $R_i$ was the net reaction (nmol $cm^{-3}$ $d^{-1}$) for the $z^{th}$ depth, $Z_{i,min}$ was the minimum depth considered for the $i^{th}$ species, and $Z_{i,max}$ was the maximum depth considered for the $i^{th}$ species. The maximum depth was set to 11cm, 40 cm, and 62 cm for TOC, $CH_4$, and $SO_4^{2-}$, respectively. The minimum depth was set to 0 cm for TOC, $CH_4$, and $SO_4^{2-}$. For both $CH_4$ and $SO_4^{2-}$, $R_i$ was taken as the solution to the reaction-transport model (equation 4.10.2). The net reaction, $R_i$, for TOC was treated as the TOC oxidation rate constant between depths 5 cm and 11 cm. This was determined by taking a linear regression of TOC loss as a function of time. Time was determined

from depth by dividing $\Delta z$ by burial rate ($\omega$=0.31 cm $yr^{-1}$). The net sinks for $CH_4$ and $SO_4^{2-}$ were determined for all values of groundwater upwelling evaluated in equation 4.10.2.

**Castle Hayne Aquifer Sulfate Concentrations**

I evaluated for $Br^-$, $Ca^{2+}$, $Cl^-$, $K^+$, $Mg^{2+}$, $Na^+$, and $SO_4^{2-}$ concentrations in the Castle Hayne Aquifer using water quality data available from the Water Quality Portal (https://www.waterqualitydata.us/). Wells sites were limited to the Castle Hayne Limestone aquifer, "accepted" concentration measurements, and well sites in Onslow County (site of sample site) and all adjacent counties (Pender, Carteret, Jones, Duplin). When concentrations were below detection, concentrations were treated as 0 mM for statistical purposes.

## Results

**Porewater Ion Depth Profiles**

The detection limits for $PO_4^{3-}$, $NH_4^+$, $Ca^{2+}$, $K^+$, $Na^+$, $Mg^{2+}$, $Cl^-$, $Br^-$, and $SO_4^{2-}$ were 0.02 mM, 0.04 mM, 0.45 mM, 0.1 mM, 7.0 mM, 0.8 mM, 12.0 mM, 0.03 mM, and 0.2 mM, respectively. The porewater ion concentrations for $PO_4^{3-}$, $NH_4^+$, $Ca^{2+}$, $K^+$, $Na^+$, $Mg^{2+}$, $Cl^-$, $Br^-$, and $SO_4^{2-}$ concentrations from the anion/cation and $SO_4^{2-}$ aliquots spanned n.d.–0.11 mM, 0.12–0.96 mM, 2.3–8.6 mM, 2.9–6.5 mM, 103.3–298.9 mM, 6.6–31.4 mM, 153.8–327.9 mM, 0.01–0.49 mM, and n.d.–17.2 mM, respectively (Figure 4.3A-I). Porewater ion concentrations for $Ca^{2+}$, $K^+$, $Na^+$, $Mg^{2+}$, $Cl^-$, $Br^-$, and $SO_4^{2-}$ followed similar trends through the sediment core, where concentrations were highest at the sediment-water interface (SWI) and decreased with increasing depth. At ~15 cmbsf, all porewater ions gradients became less steep. Taking a piecewise linear regression with a node at 15 cmbsf, I calculated gradient changes for $Ca^{2+}$, $K^+$, $Na^+$, $Mg^{2+}$, $Cl^-$, and $Br^-$. Slopes are reported in Table 4.1. In contrast, $PO_4^{3-}$ and $NH_4^+$ concentrations were lowest

at the sediment-water interface (SWI) and increased with depth. The $PO_4^{3-}$ concentrations reached an asymptote after 10 cmbsf while $NH_4^+$ concentrations increased with depth to the bottom of the collected core.

To ensure changes in concentration gradients were not artifacts caused by a change in the angle of the core as it was pushed into the sediment, I calculated the theoretical change in sampling angle that would result in the observed changes in concentration profile if the true profile were linear. The bias in sampling angle necessary to change the slope between 3 to 15 cmbsf and 16 to 60 cmbsf for $Br^-$, $Ca^+$, $Cl^-$, $K^+$, $Mg^{2+}$, and $Na^+$ were 87.4° 85.2° 83.7° 86.6° 84.0°, and 84.5°, respectively. Considering all seawater ions (excluding $SO_4^{2-}$ due to $SO_4^{2-}$ reduction), I estimated it would take an 85.2±1.8° change in sampling angle to explain the change in concentration gradients. Although it is difficult to precisely push a core vertically into sediment, I judge that this amount of change in the angle is unrealistically large, and do not consider this possibility further.

**Castle Hayne Aquifer Ion Concentrations**

Geological surveys indicate he Castle Hayne Aquifer underlies the White Oak River Estuary (139). I performed a regional analysis of $Ca^{2+}$, $K^+$, $Na^+$, $Mg^{2+}$, $Cl^-$, $Br^-$, and $SO_4^{2-}$ concentrations in the Castle Hayne Aquifer using data available from the Water Quality Portal. The average concentrations (± standard deviation) for $Ca^{2+}$, $K^+$, $Na^+$, $Mg^{2+}$, $Cl^-$, $Br^-$, and $SO_4^{2-}$ was 2.04±1.51 mM (*n*=26), 0.07±0.06 mM (*n*=26), 0.56±1.24 mM (*n*=26), 0.12±0.10 mM (*n*=26), 0.31±0.27 mM (*n*=26), $7.8\times10^{-4}$ $\pm1.22\times10^{-3}$ mM (*n*=19), and 0.08±0.25 mM (*n*=26), respectively.

**Organic Carbon Depth Profiles**

Dissolved $CH_4$ concentrations followed a sigmodal profile as a function of depth (Figure 3A). The lowest dissolved $CH_4$ concentrations (~0 mM) occurred at the SWI, and the maximum concentration (1.91 mM) occurred at ~45 cmbsf. After reaching the maximum, dissolved $CH_4$ stabilized at ~1 mM for depths >50 cmbsf.

Values for $\delta^{13}C_{organic}$ from the sediment changed from -26.04 ‰ near the SWI to -23.40 ‰ near the bottom of the sediment core. This transition included an average $\delta^{13}C_{organic}$ change of 0.20‰ $cm^{-1}$ ($p$-value<$10^{-3}$) from 4 to 10 cm, which did not vary between 10 and 30 cmbsf, but shifted to -0.11‰ $cm^{-1}$ ($p$-value<$10^{-3}$) from 30 to 35 cmbsf, and 0.06‰ $cm^{-1}$ ($p$-value<$10^{-7}$) between 35 to 58 cmbsf.

TOC and TON profiles from the sediment covaried with depth. TOC varied from 1.39 to 2.99 mmol $cm^{-3}$ (Figure 3C). A TOC maximum occurred at 3–4 cmbsf, then decreased at 10 cmbsf until 2.25 mmol $cm^{-3}$ at 32 cm and then to 1.65 mmol $cm^{-3}$ at 62 cmbsf. TON varied from 0.08 to 0.17 mmol $cm^{-3}$(Figure 3D), with a maximum at 3–4 cmbsf, minimum at 10 cmbsf, then another peak of 0.11 mmol $cm^{-3}$ at 32 cm, and decrease to 1.65 mmol $cm^{-3}$ at 62 cmbsf. An oscillatory pattern in TOC:TON ratios (C:N) occurred through the sediment core (Figure 3E). C:N values fluctuated between 15.2 and 18.3, with local maxima occurring at 3, 10, 40, and 62 cmbsf and local minima occurring at 8, 28, and 55 cmbsf.

DOM spectral slope of the extracted pore water varied from 2.5 to 8.5 (Figure 3F). The spectral slope maximum occurred at 3 cmbsf and was followed by a decrease with depth until ~10 cmbsf. The spectral slope became invariant with at depths >10 cmbsf.

**Groundwater Upwelling and Sulfate/Methane Production Rates**

A reaction-transport model was used to infer oxidation, reduction, and production rates for $SO_4^{2-}$ and dissolved $CH_4$ (equations 4.7-4.10). The GAM smooth fits used for the $SO_4^{2-}$ and dissolved $CH_4$ concentration gradients ($\frac{\partial C_i}{\partial z}$) had deviance explained values ($p$-value<$10^{-4}$) of 96.6% and 82.9%, respectively. (Figure 4.5A-B). The lower boundary condition for $SO_4^{2-}$ concentrations predicted from the GAM (0.14 mM) was comparable to $SO_4^{2-}$ concentration reported in the Water Quality Portal (see above).

When groundwater upwelling was 0 cm $yr^{-1}$, net $SO_4^{2-}$ production was negative at all depths of the sediment column (Figure 4.6A). The net $SO_4^{2-}$ production rates followed a modal profile with depth. The minimum net $SO_4^{2-}$ production was -35.6 nM $SO_4^{2-}$ $d^{-1}$ and occurred at 15 cmbsf. The net $SO_4^{2-}$ production rates converge towards 0 nM $SO_4^{2-}$ $d^{-1}$ near the SWI and 50 cmbsf. When I increased the groundwater upwelling rate, net $SO_4^{2-}$ production increased at depths from 0 to ~40 cmbsf. At some depths, negative net $SO_4^{2-}$ production transitioned from a negative production rate to a positive production. The depth that positive $SO_4^{2-}$ production rates occurred increased as a function of groundwater upwelling rates. The deepest depths where positive $SO_4^{2-}$ production occurred were 4, 5, 8, 10, 12, and 14 cmbsf when groundwater upwelling rates were 1, 2, 5, 10, 15, and 20 cm $yr^{-1}$, respectively.

In contrast to net $SO_4^{2-}$ production when groundwater upwelling was 0 cm $yr^{-1}$, $CH_4$ production and oxidation occurred simultaneously at different depths (Figure 4.6B). Dissolved $CH_4$ oxidation followed a modal distribution between 0 and 38 cmbsf. The maximum dissolved $CH_4$ oxidation rate was -4.3 nM $d^{-1}$ at 32 cmbsf. Dissolved $CH_4$ production also followed a modal distribution and spanned between 38 and 60 cmbsf. The maximum dissolved $CH_4$

production rate was 6.5 nM $CH_4$ $d^{-1}$ and occurred at 47 cmbsf. An increase in groundwater upwelling had minor effects on depths of the maximum and minimums, as well as the general distributions for $CH_4$ production and oxidation. Nonetheless, groundwater upwelling oppositely influenced the magnitudes of the absolute maximum and minimums for $CH_4$ production and oxidation rates. The maximum dissolved $CH_4$ oxidation rates were -4.3, -4.5, -4.9, -5.5, -6.0, and -6.7 nM $CH_4$ $d^{-1}$ when groundwater upwelling rates were 1, 2, 5, 10, 15, and 20 cm $yr^{-1}$, respectively. The maximum dissolved $CH_4$ production rates were 6.5, 6.4, 6.2, 5.8, 5.4, and 5.1 nM $CH_4$ $d^{-1}$ when groundwater upwelling rates were 1, 2, 5, 10, 15, and 20 cm $yr^{-1}$, respectively.

**Carbon and Sulfur Budgets in Station H White Oak River Estuary Sediments**

I evaluated TOC, DIC and $SO_4^{2-}$ reservoirs in WOR estuary, Station H sediments. All flux estimates reported here can be found in Table 4.2. We observed two notable features in the TOC profiles at Station H. There was a net loss of 1.22 mmol TOC $cm^{-3}$ spanning 0 to 10 cmbsf. This loss was proceeded by a modal increase in TOC, spanning from 10 to 60 cmbsf, with a maximum concentration of 2.20 mmol TOC $cm^{-3}$ at 35 cmbsf. The net change in TOC from 10 to 60 cmbsf was ~0 mmol TOC $cm^{-3}$ (Figure 4.4C). I converted sediment depth to units of time by dividing depth intervals by an assumed burial rate (ω) of 0.31 cm $yr^{-1}$ (34, 130, 131). I regressed TOC loss from 0 to 10 cmbsf as a function of time (mmol TOC $cm^{-3}$ $yr^{-1}$), with the exclusion depths 3 and 4 cmbsf. This resulted in an estimated average TOC loss of -37.9 mmol TOC $cm^{-3}$ $yr^{-1}$ (Figure 4.7). As, the net change from 10 to 60 cmbsf was ~0 mmol TOC $cm^{-3}$ (Figure 4.4C), I attributed all TOC losses in the first ~62 cm to TOC occurring between 0 and 10 cmbsf. Thus, assuming a constant rate of TOC loss in the first 10 cmbsf (Figure 4.7), the integrated annual loss in TOC was estimated 3178 mmol $m^{-2}$ $yr^{-1}$. As $CH_4$ production is negligible (see below) relative

to TOC concentrations, I take TOC loss as being equivalent to DIC flux (130). Last, TOC burial flux was taken as the difference between TOC depositional flux at the SWI and the DIC flux. Assuming a burial rate of 0.31 cm $yr^{-1}$, and TOC concentration of 2.61 mmol TOC $cm^{-3}$ at 0 cmbsf, I estimated an OC depositional flux of 8.09 mmol $m^{-2}$ $yr^{-1}$ and a TOC burial flux of 4.91 mmol $m^{-2}$ $yr^{-1}$.

The $CH_4$ and $SO_4^{2-}$ production rates varied as a function depth and groundwater upwelling rates (Figure 4.6). Thus, I quantified column integrated net $CH_4$ and $SO_4^{2-}$ production as a function of groundwater upwelling rates. Column integrated $CH_4$ flux was inversely related to groundwater upwelling rates (Figure 4.8). When groundwater upwelling was 0 cm $yr^{-1}$, WOR sediments acted as a net source of $CH_4$, with a $CH_4$ production flux of 34.7 mmol $CH_4$ $m^{-2}$ $yr^{-1}$. However, when groundwater upwelling exceeded ~3 cm $yr^{-1}$, WOR sediments became a net sink for $CH_4$. The maximum evaluated flux ($u$=20 cm $yr^{-1}$) resulted in a net $CH_4$ flux of -182.6 mmol $CH_4$ $m^{-2}$ $yr^{-1}$. In contrast to $CH_4$, column integrated $SO_4^{2-}$ production flux positively related to groundwater upwelling rates (Figure 4.9). When groundwater upwelling was 0 cm $yr^{-1}$, the WOR sediments acted as a sink for $SO_4^{2-}$, with a $SO_4^{2-}$ flux of -1976 mmol $SO_4^{2-}$ $m^{-2}$ $yr^{-1}$. However, when groundwater upwelling exceeded ~15 cm $yr^{-1}$, WOR sediments became a net source for $SO_4^{2-}$. The maximum evaluated $SO_4^{2-}$ flux ($u$=20 cm $yr^{-1}$) was 672 mmol $SO_4^{2-}$ $m^{-2}$ $yr^{-1}$.

## Discussion

Here, I sampled a 62 cm long sediment core from the WOR estuary, Station H, with the intention of characterizing gradients in microbial traits with respect to geochemical gradients (135). Site accessibility, a large body of literature spanning several decades, and a diverse suite of redox metabolisms (associated electron acceptors) makes the WOR estuary a desirable model

system for modeling microbial trait distributions with respect to geochemistry (17, 34, 129, 131, 132, 134, 142). Prior to exploring possible ecological niches at the WOR, I sought to evaluate inventories of TOC, DIC, $CH_4$, and $SO_4^{2-}$, the primary electron donors (TOC and $CH_4$) and acceptor (DIC and $SO_4^{2-}$) in WOR sediments. As a part of this study, I quantified major seawater cations ($Na^+$, $Mg^{2+}$, $K^+$, $NH_4^+$, and $Ca^{2+}$), anions ($SO_4^{2-}$, $Cl^-$, $PO_4^{3-}$, $Br^-$), and DOM spectral slope from porewater as well as TOC, TON, bulk organic C isotope composition, and dissolved $CH_4$ from bulk sediment (Figure 4.2).

**Castle Hayne Aquifer and Ion Dilution in WOR Sediments**

The decrease in major seawater ion concentration with sediment depth was previously reported in WOR sediments (129, 133). The reduction in concentrations can be explained by intrusion of groundwater from the underlying Castle Hayne aquifer. Geological surveying indicates that the confining unit for the Castle Hayne aquifer is thin (~4.2 m) and porous across Eastern North Carolina. In some instances, the confining unit has been reported to be absent, leading to direct contact between river/estuary systems and the Castle Hayne aquifer (133, 138, 139). Here, porewater ion concentrations decreased with depth for all major seawater ions (Figure 4.3). The rate of decrease between 3 and 15 cmbsf and 16 and 60 cmbsf changed for all major seawater ions (Table 4.1). At present, it is unclear whether the change in rates reflects a lag in diffusion from oscillating water column salinity, the presence of fresh groundwater upwelling, or both. High frequency variability in water column salinity would give rise to stable ion profiles at depth that could be used for inferring groundwater upwelling rates (145). Previous studies assumed steady-state thermal gradients (138) and transport modeling with $Cl^-$ concentrations (treated as a conservative tracer) (133) to infer groundwater upwelling. In these

studies, groundwater advection estimates spanned from 0.6 to 15.3 cm $yr^{-1}$ depending on the time of year at Station H. My analysis of groundwater ion concentrations indicated discharge from the underlying Castle Hayne Aquifer, the underlying aquifer, would lead to a decrease in porewater ion concentrations, as all ion concentrations are orders of magnitude less than seawater ion concentrations. The relative importance of the advection versus diffusion terms (equation 4.7) in explaining the dissolved ion profiles are difficult to ascertain without salinity timeseries data from the WOR water column.

**Metabolic Alterations to Organic Carbon in WOR Sediments**

Sediments in the United States Coastal Plains estuaries are characterized by high OC burial with low oxygen penetration (127). The WOR sediments analyzed here were no exception, with OC burial efficiency of ~60% (Figure 4.4C). Although debates still exist about the primary variable controlling OC turnover in sediments, porewater $O_2$ concentration and exposure time is considered one of the more likely candidates influencing OC turnover (126, 146). High carbon burial rates correlate with carbon burial efficiency as a result of higher aerobic respiration rates depleting oxygen at shallow depths (147). Although I did not directly measure $O_2$, I estimate that $SO_4^{2-}$ reduction initiated at ~5 cmbsf (based on the first derivatives of normalized seawater concentrations). This observation implied microbial metabolisms utilizing aerobic and sub-oxic metabolism types were depleted in the top 5 cmbsf (17, 32). Note that complete titration of $SO_4^{2-}$ prior to loss in other ion concentrations indicates the presence of sulfate-reducing metabolisms. Once $SO_4^{2-}$ titrated to 0 mM, methanogenesis became the dominant source of energy in WOR sediments; this was evident by the increase from ~ 0 mM to 1 mM $CH_4$ $cm^{-3}$ between 35 and 50

cmbsf. Generally, these concentrations were comparable with previously reported $CH_4$ concentrations in WOR sediments (131, 133, 134).

One feature in the OM profiles was an increase in TOC and TON concentrations at 35 cmbsf (Figure 4.4C,D). To our knowledge, this is the first observation of this pattern in WOR sediments from Station H (34, 35, 129, 130). As this maximum is located at the SMTZ (Figure 4.3I and Figure 4.4A), I hypothesize the increase in OM is from biomass production generated from anaerobic oxidation of methane (AOM) (137). Anaerobic methanotrophic archaea (ANME), which perform AOM, have been observed in WOR sediments at the SMTZ (131, 134). The primary evidence supporting this hypothesis is the local minimum in $\delta^{13}C_{organic}$ (Figure 4.4B) and local maximum in C:N (Figure 4.4E), which coincides with both the SMTZ and the local OM maximum at 35 cmbsf. Biomass production from oxidized $CH_4$ should result in OC with a lighter isotopic signature as carbon from biogenic $CH_4$ is isotopically light ($\delta^{13}C < -58‰$) compared to other organic molecules ($\delta^{13}C > -30‰$) (148).

In addition to the prominence of sulfate-reducers and methanogens, fermenters are prevalent in WOR sediments (17). Fermentative metabolisms are thought to transform organic polymers into respiration products which include $CO_2$, $H_2$, organic acids (e.g., acetate), $NH_4^+$, $PO_4^{3-}$, and alcohols (e.g., ethanol) (149). The ecological role of fermenters is significant as respiration products provide $H_2$, a source of electrons, for methanogens and sulfate-reducers. Here, I observed indirect evidence of fermentative metabolisms via the elevated $NH_4^+$ and $PO_4^{3-}$ concentrations (Figure 4.3G,H). Contrary to major seawater ion concentration profiles, the $NH_4^+$ and $PO_4^{3-}$ concentrations were lowest near the SWI and increased with depth. Changes in $NH_4^+$ and $PO_4^{3-}$ stoichiometric ratios with depth may indicate a change in OM quality (10 to 30) if

depositional sources OM did not change with time. Lastly, the $S_R$ decrease suggests a drop in fulvic to humic acid ratios (141) (Figure 4.4I). The coinciding $S_R$ decrease at depths where $SO_4^{2-}$ reduction begins suggests the drop in fulvic acids may be from fermentative metabolisms, which operate in anaerobic environments (149).

**Evaluating the Carbon and Sulfur Budgets at WOR, Station H**

In WOR sediments below 5 cmbsf, the dominant electron donors and acceptors are carbon and sulfur, respectively. For this reason, fluxes for major carbon and sulfur reservoirs in WOR sediments were estimated via concentration measurements (as well as acoustic turbidity for $CH_4$) and reaction-transport modeling (34, 35, 130, 131). I estimated TOC, DIC, $CH_4$, and $SO_4^{2-}$ fluxes based on the measurements (Figure 4.3-4.4) presented here and reactive-transport modeling (Figure 4.5-4.6). As TOC chemical transport is controlled solely by sediment burial, TOC burial flux and DIC flux estimates were invariant to simulations which consider groundwater upwelling (Table 4.1). DIC fluxes should be sensitive to advective processes as DIC exists as dissolved $HCO_3^-$ for pH values typical of WOR sediments (129). Nonetheless, our DIC flux estimates, based on oxidized TOC, were comparable with flux measurements derived from DIC gradient analysis (130).

When assuming diffusion-dominated chemical transport, our $CH_4$ and $SO_4^{2-}$ flux estimates were elevated but generally comparable with previous estimates (Table 4.1). The elevated $CH_4$ and $SO_4^{2-}$ fluxes can be explained by steeper concentration gradients compared to gradients used in earlier modeling studies (34, 35). $SO_4^{2-}$ removal would suggest WOR sediments act as a sink for $SO_4^{2-}$ from the overlying water column. Here, $SO_4^{2-}$ removal is likely driven through $SO_4^{2-}$ reducing metabolisms (17). Excess $CH_4$ is assumed to diffuse away from

the SWI, where the $CH_4$ flux is 0 mmol $m^{-2}$ $yr^{-1}$. Acoustic turbidity measurements suggest $CH_4$ penetration ceases at ~1 m (34, 133). Here, I did not collect sediment to this depth and cannot heavily speculate about $CH_4$ sink processes below 62 cmbsf.

Earlier studies on thermal gradients (138) and $Cl^-$ concentration profiles (assumed as passive tracer) (133) estimate groundwater upwelling rates between 0.6 to 39.4 cm $yr^{-1}$ in WOR sediments. Nonetheless, $CH_4$ and $SO_4^{2-}$ flux modeling studies for WOR sediments have not considered groundwater upwelling (34, 35, 130, 131). I found that net $CH_4$ flux decreased and net $SO_4^{2-}$ flux increased with increasing advection (Figure 4.8-4.9; Table 4.2). The decreases in net $CH_4$ flux can be explained by the reaction term (equation 4.7) compensating for $CH_4$ advecting from lower depths (Figure 4.6B). Net $CH_4$ flux becomes negative when advection >3.2 cm $yr^{-1}$. Since all dissolved $CH_4$ is produced via methanogenesis in WOR sediments (130, 133), dissolved $CH_4$ production must happen below 62 cmbsf to maintain mass balance if advection rates are >3.2 cm $yr^{-1}$. The increase in net $SO_4$ flux is explained by less $SO_4^{2-}$ supply from the overlying water column due to advection. The consequence of less $SO_4^{2-}$ penetration was the reaction term became positive, indicating $SO_4^{2-}$ production was necessary to maintain steady-state $SO_4^{2-}$ concentrations (Figure 4.6A). As advection rates increased, the $SO_4^{2-}$ production was predicted to occur at deeper depths and at larger magnitudes. The sediment sulfur cycle has many cryptic (short lived or below detection) sulfur species of intermediate oxidation state (150). The oxidation of reduced and intermediate sulfur species can be driven by biotically and abiotically. Notably, disproportionation reactions with intermediate oxidation state sulfur species generates both $SO_4^{2-}$ and $H_2S$, Fe and Mn can act as oxidants, and bioirragation can lead to localized microaerophilic zones in sediments. Metabolisms involved with sulfur species of intermediate

oxidation states have been observed in WOR sediments (17). Furthermore, porewater sulfide has been observed to extend to the SWI (134). It is conceivable that reduced sulfur species are respired at depth, transported to depths where $SO_4^{2-}$ production is predicted with our reaction-transport model, and then re-oxidized back to $SO_4^{2-}$. Kinetically, which processes would support the production of $SO_4^{2-}$ observed in the model is unclear and requires further investigation (150). Ultimately, early $SO_4^{2-}$ sequestration estimates for WOR estuary sediments appear overestimated if groundwater advection is considered in models. Furthermore, short and intermediate variability in groundwater advection rates may lead to fluctuations in WOR sediments acting as a source or sink for $SO_4^{2-}$ (Figure 4.9).

**Conclusions and Future Directions**

We sought to characterize bulk sediment and porewater geochemistry prior to modeling microbial trait distributions in WOR estuary sediments. The TOC, DIC, $CH_4$, and $SO_4^{2-}$ budgets for primary carbon and sulfur reservoirs are consistent with previous estimates; however, variability in the SMTZ between different studies (34, 35, 129, 131, 133, 134) indicate that assumptions of steady-state concentrations on short and intermediate timescales may be unadvisable when characterizing microbial ecology. AOM may result in biomass growth at the SMTZ (Figure 4.4). Furthermore, sulfur cycling appeared sensitive to groundwater upwelling (Figure 4.6A), which, if variable, could have implications on temporal microbial populations.

The geochemical results was consistent with the expected metabolisms previously described in WOR sediments (17, 131, 132, 134). Evaluation of microbiology and via genomic sequencing and proteomics may prove useful in validating covariance in these metabolism types with observed trends in geochemical parameters. However, the extent that these observations

will be generalizable, given the short and intermediate variability of geochemistry in WOR sediments, will remain under question. Thus, future work should focus on understanding which variables influence short and intermediate timescale (i.e., days to months) variability in geochemistry. I view this as important moving forward with modeling microbial ecology as it is unclear how buffered the microbial ecology is with respect to stochastic variability in geochemistry. Variables that I think may be important on day to month timescales include changes in groundwater upwelling rates, OM depositional fluxes, and water column salinity. I explored using $Cl^-$ concentrations in a transport model (along with other seawater ions) to determine groundwater advection with equation 4.7 (133). Groundwater upwelling estimates were on average ~20 cm $yr^{-1}$ using this approach and exceeded the 15 cm $yr^{-1}$ constraint based on $SO_4^{2-}$ mass balance. Thus, characterizing groundwater upwelling rates may need a better proxy than seawater ions. One candidate would be $SF_6$, an inert gas. An experimental design may include inoculating $SF_6$ into WOR sediments and monitoring vertical transport over time.

# Appendix Four

## Tables

**Table 4.1.** Pairwise slopes for porewater ion concentrations. All slopes are significant ($p$<0.001).

| Depth (cmbsf) | $Br^-$ (mM $cm^{-1}$) | $Ca^{2+}$ (mM $cm^{-1}$) | $Cl^-$ (mM $cm^{-1}$) | $K^+$ (mM $cm^{-1}$) | $Mg^{2+}$ (mM $cm^{-1}$) | $Na^+$ (mM $cm^{-1}$) |
|---|---|---|---|---|---|---|
| 3–15 | -0.02 | -0.36 | -11.1 | -0.24 | -1.47 | -12.1 |
| 16–60 | -0.001 | -0.02 | -1.2 | -0.01 | -0.15 | -1.1 |

**Table 4.2.** A summary of carbon and sulfur budgets in sediments at WOR estuary, Station H.

| Study | Groundwater Upwelling[ǂ] | $SO_4^{2-}$ Removal[*] | $SO_4^{2-}$ Production[*] | $CH_4$ Oxidation[*] | $CH_4$ Production[*] | TOC Burial[*] | DIC Production[*] |
|---|---|---|---|---|---|---|---|
| (34, 35) | 0 | 1300 | -- | 145 | 196 | -- | -- |
| (130) | 0 | -- | -- | -- | -- | 5259 | 4032 |
| Here | 0 | 1995 | 18 | 239 | 274 | 6091 | 3178 |
| | 14.9 | 739 | 739 | 344 | 217 | 6091 | 3178 |

ǂ Units of cm $yr^{-1}$

* Units of mmol $m^{-2}$ $yr^{-1}$

**Figures**


**Figure 4.1.** A map (Google Earth) showing the White Oak River Estuary and Station H.

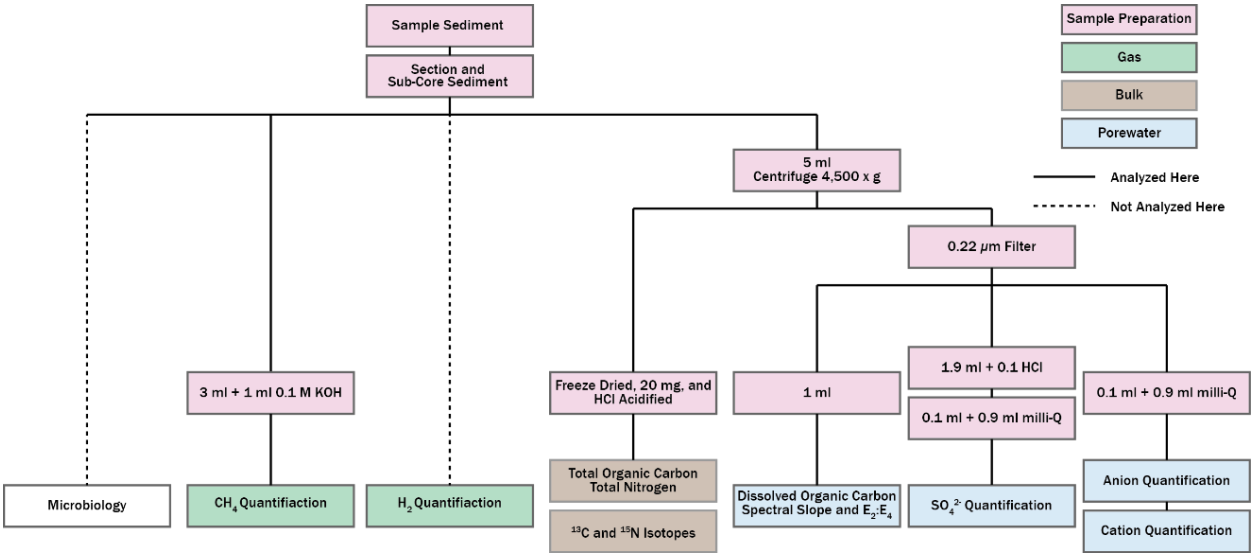


**Figure 4.2.** A flow diagram showing sample processing and analyzes performed on sediments.

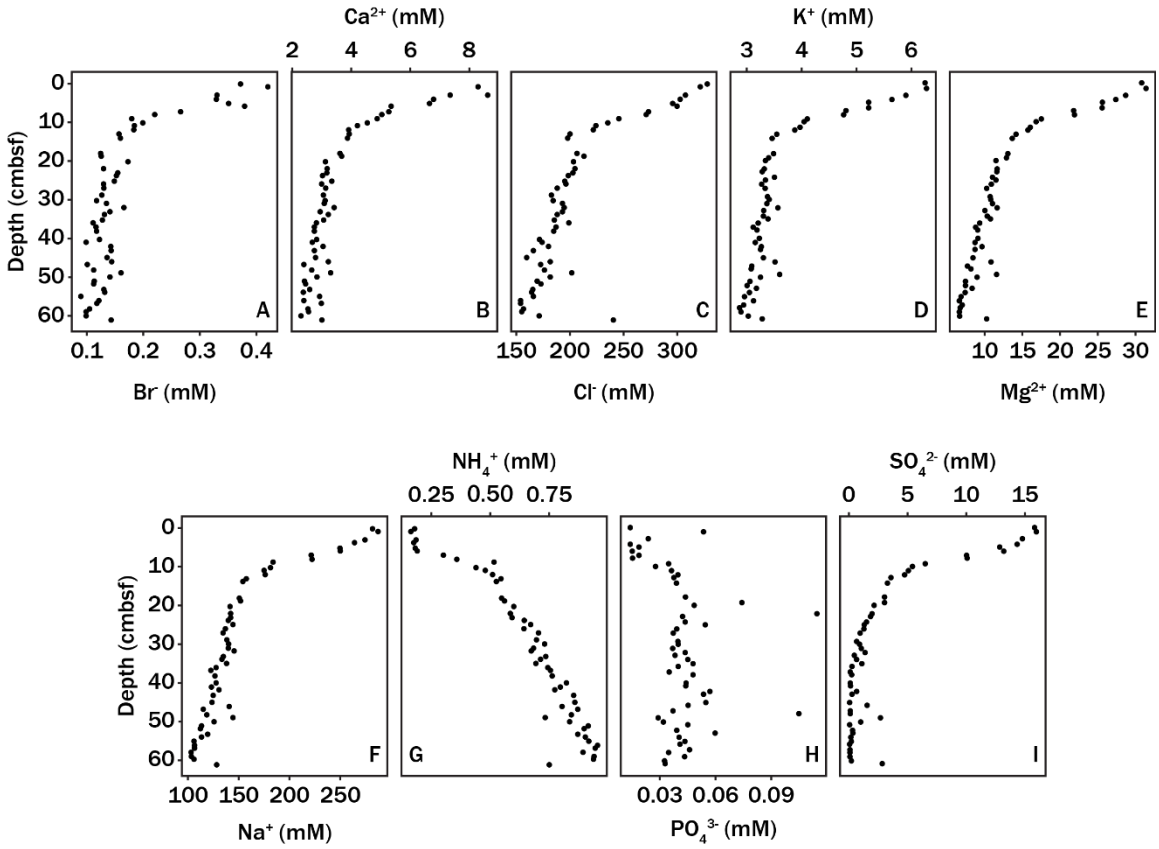


**Figure 4.3.** Porewater concentration profiles (mM) for $Br^-$ (A), $Br^-$ (A), $Ca^{2+}$ (B), $Cl^-$ (C), $K^+$ (D), $Mg^{2+}$ (E), $Na^+$ (F), $NH_4^+$ (G), $PO_4^{3-}$ (H), and $SO_4^{2-}$ (I).

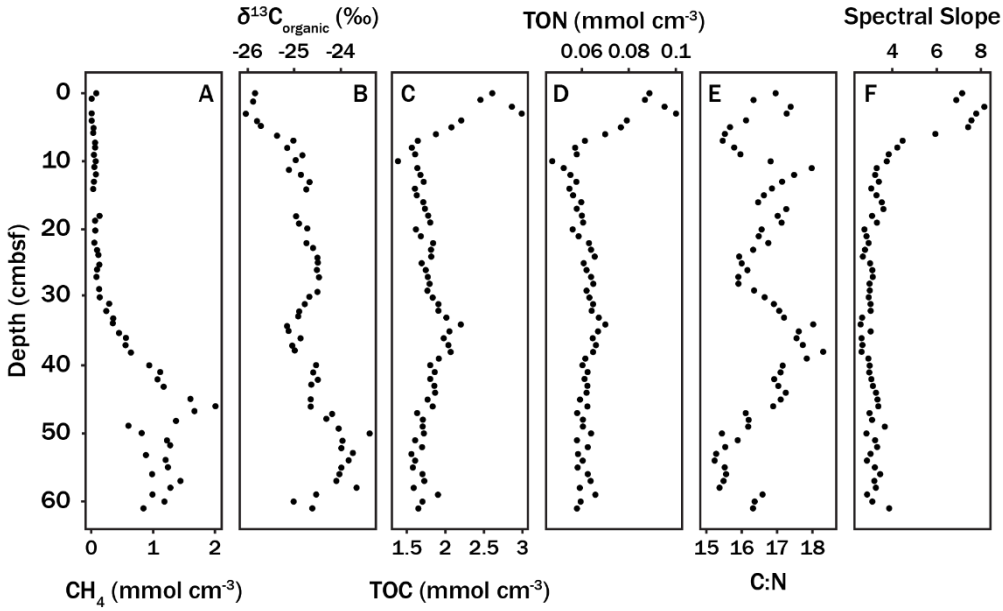


**Figure 4.4.** Organic carbon concentration profiles for $CH_4$ (mmol $cm^{-3}$) (A), bulk $\delta^{13}C_{organic}$ (‰ with respect to VPDB) (B), TOC (mmol $cm^{-3}$) (C), TON (mmol $cm^{-3}$) (D), carbon to nitrogen ratio (E), and DOM spectral slope (F).

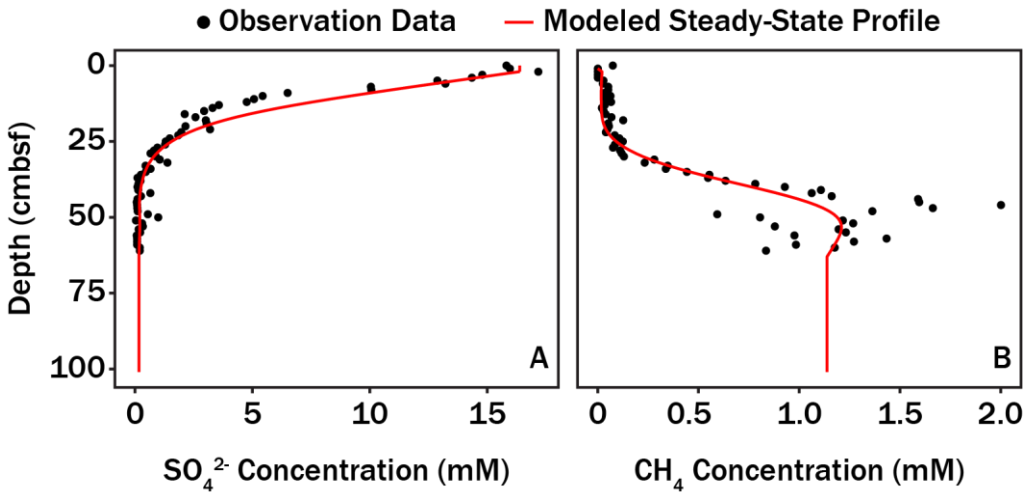


**Figure 4.5.** Generalized additive model (GAM) fits for $SO_4^{2-}$ depth profiles (mM) (A) and dissolved $CH_4$ (mM) (B) used for reaction-transport modeling.

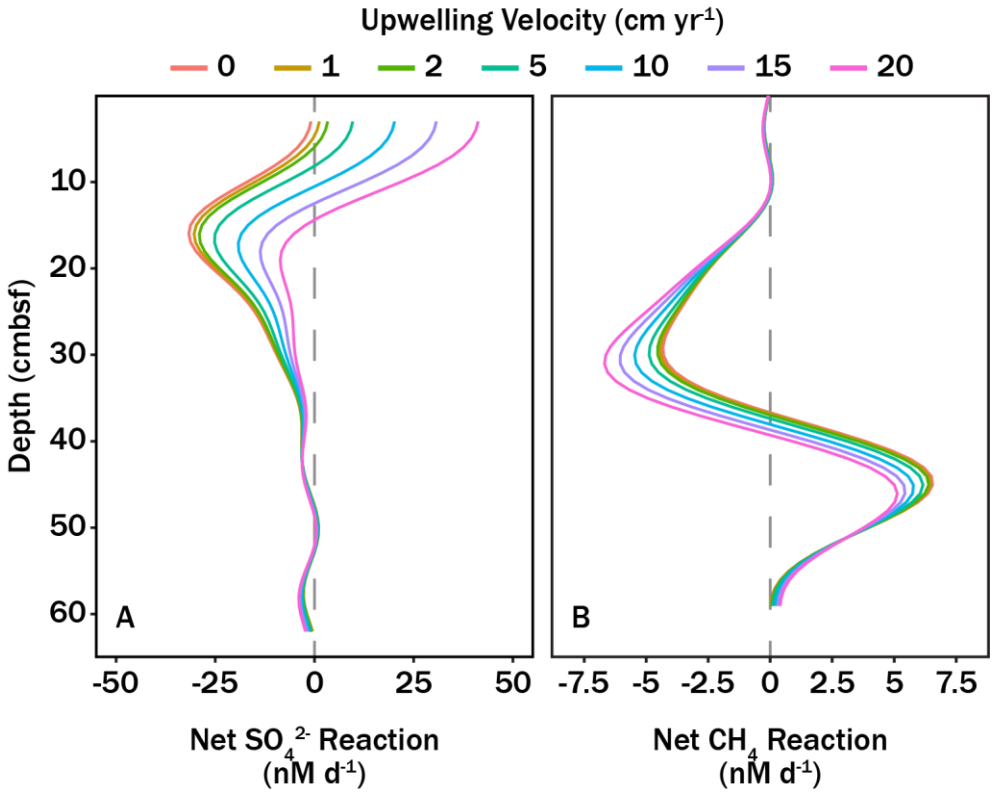


**Figure 4.6.** Model results for $SO_4^{2-}$ (A) and dissolved $CH_4$ (B) from the reaction-transport model. Contours correspond to different assumed rates (cm $yr^{-1}$) of groundwater upwelling.

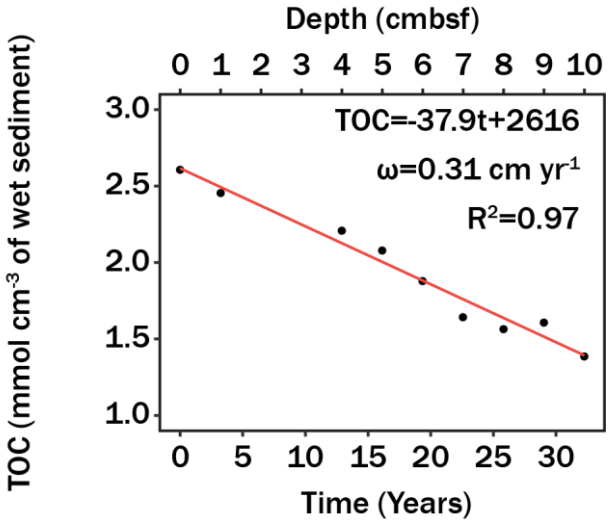


**Figure 4.7.** A linear regression fit of TOC loss as a function of time assuming an average sedimentation rate of 0.31 cm $yr^{-1}$ (34, 130, 131).

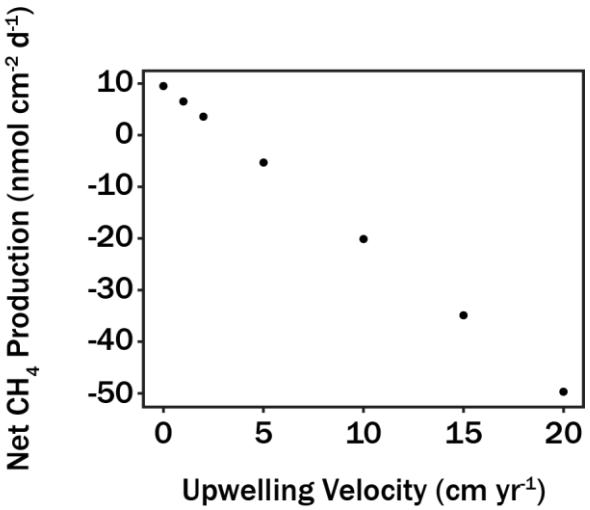


**Figure 4.8.** Column integrated $CH_4$ production (nmol $cm^{-2}$ $d^{-1}$) predicted from the reaction-transport model (Figure 4.4B) as a function of groundwater upwelling (cm $yr^{-1}$).

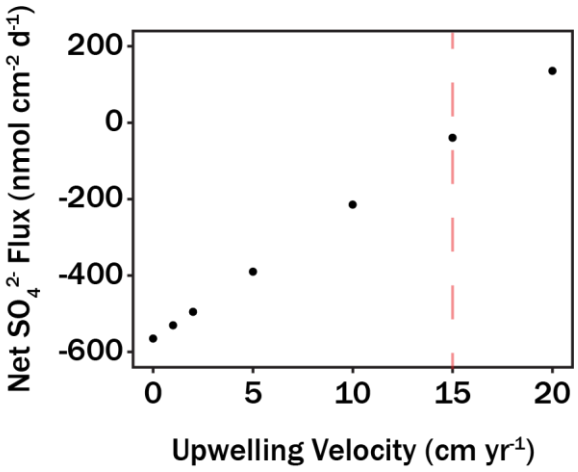


**Figure 4.9.** Column integrated $SO_4^{2-}$ production (nmol $cm^{-2}$ $d^{-1}$) predicted from the reaction-transport model (Figure 4.4A) as a function of groundwater upwelling (cm $yr^{-1}$). The dashed red line corresponds to the groundwater upwelling velocity where WOR sediments at Station H become a source for $SO_4^{2-}$.

# CONCLUSIONS

It seems unlikely that a reasonable constraint on the number of unique metabolisms or the spatiotemporal distribution of those metabolisms is to happen soon. Again, new metabolisms are theorized and discovered on a regular basis (5) and a large fraction of genes found in metagenomes remain unannotated (19). Nonetheless, the work presented here made strides towards addressing these questions. Here, theoretical and quantitative approaches were explored for characterizing microbial metabolism diversity from several different perspectives. First, microbial metabolism diversity was evaluated with respect to phylogenetics, or evolutionary relatedness, for lineages across the microbial tree of life. Metabolic coherence was observed in taxonomic ranks spanning phylum to genus when annotating traits with cluster of orthologous groups (COGs). Higher taxonomic ranks (phylum, class, order) had the greatest explanatory power towards explaining metabolic potential and suggests broad-scale patterns may exist for specific lineages occupying ecological niches. Both, the average fraction of genomic content derived via vertical gene transfer and taxonomy's predictive power towards genomic content is now quantitatively evident. Improving annotation rates for more metabolically-detailed annotation schemes, such as KEGG, is a viable path moving forward as these models would provide predictions related to specific metabolites. Second, a theoretical framework was proposed for prescribing a sequencing effort in shotgun metagenomic sequencing experiments. The theoretical model linked genome rareness to an expected short read sequencing effort necessary to meet a genome coverage criterion defined by an investigator. This model was then evaluated with respect to MAG retrieval from four different sample sites: maize soil, surface ocean, human gut, and estuary sediments. Normalizing the sequencing effort applied in

sequencing experiments to a target genomic content rareness should improve comparability among studies with respect to MAG discovery and metabolism discovery rates. Moving forward, continued improvements in bioinformatic pipelines with respect to assembly and binning algorithms should reduce the sequencing coverage and overall sequencing effort necessary to retrieve MAGs. Third, a theoretical framework was proposed for measuring trait functional redundancy within microbial communities. The functional redundancy metric was modeled after traditional diversity theory and was referred to as trait contribution evenness (TCE). Functional redundancy among different nitrogen-cycling pathways in *TARA* Oceans MAGs was then analyzed as a case study. TCE provides a measure for spatiotemporal variations of environmental trait distributions. This measure will aid in addressing broader questions about the ecological significance of functional redundancy in microbial communities and trait-based modeling. Lastly, organic geochemistry was measured in an estuary sediment core that covered three redox conditions: oxic (and sub-oxic), $SO_4^{2-}$ rich (anoxic), and $CH_4$ rich (anoxic). Carbon and sulfur budgets were modeled from organic geochemistry and were subsequently related to microbial metabolisms known to exist in sediment environments. This work strived to establish the temporal stability in organic geochemistry as preliminary analysis for niche modeling.

As of RefSeq v92 (used in Chapter One), ~145,000 non-redundant microbial genomes have been sequence, which are classified as ~25,000 unique species (GTDB). Every genome, for every environment, typically has 1000s of genes. Properly annotating ‘omics traits, identifying important traits, determining whether these traits manifest into phenotype, and relating these traits to the environment conditions is often daunting. Even with these challenges, a lot information can be gleaned from available genomic data. A general theme from my dissertation

was developing quantitative and theoretical approaches to interpret large genomic datasets by leveraging publicly available genomic data. I view this approach as amenable for exploring ecological theory surrounding microbial metabolism diversity. Strategies are improving for integrating omics-derived traits with geochemistry (18) and in trait-based modeling (2, 27). With ongoing improvements in sequencing technology and bioinformatic pipelines, ‘omics data should only become more accessible with time (20), and in conjunction with environmental metadata, publicly available repositories should become even better for addressing broad ecological questions related to the number of unique microbial metabolisms and the spatiotemporal distribution of these metabolisms.

# LIST OF REFERENCES

1. Falkowski PG, Fenchel T, Delong EF. 2008. The Microbial Engines That Drive Earth's Biogeochemical Cycles. Science (80- ) 320:1034–1039.
2. Widder S, Allen RJ, Pfeiffer T, Curtis TP, Wiuf C, Sloan WT, Cordero OX, Brown SP, Momeni B, Shou W, Kettle H, Flint HJ, Haas AF, Laroche B, Kreft JU, Rainey PB, Freilich S, Schuster S, Milferstedt K, Van Der Meer JR, Grobkopf T, Huisman J, Free A, Picioreanu C, Quince C, Klapper I, Labarthe S, Smets BF, Wang H, Soyer OS, Allison SD, Chong J, Lagomarsino MC, Croze OA, Hamelin J, Harmand J, Hoyle R, Hwa TT, Jin Q, Johnson DR, de Lorenzo V, Mobilia M, Murphy B, Peaudecerf F, Prosser JI, Quinn RA, Ralser M, Smith AG, Steyer JP, Swainston N, Tarnita CE, Trably E, Warren PB, Wilmes P. 2016. Challenges in microbial ecology: Building predictive understanding of community function and dynamics. ISME J 10:2557–2568.
3. Kuypers MMM, Sliekers AO, Lavik G, Schmid M, Jørgensen BB, Kuenen JG, Sinninghe Damsté JS, Strous M, Jetten MSM. 2003. Anaerobic ammonium oxidation by anammox bacteria in the Black Sea. Nature 422:608–611.
4. Dalsgaard T, Canfield DE, Petersen J, Thamdrup B, Acuña-González J. 2003. N2 production by the anammox reaction in the anoxic water column of Golfo Dulce, Costa Rica. Nature 422:606–608.
5. Amend JP, Aronson HS, Macalady J, LaRowe DE. 2020. Another chemolithotrophic metabolism missing in nature: sulfur comproportionation. Environ Microbiol 00.
6. Amend JP, LaRowe DE. 2019. Minireview: demystifying microbial reaction energetics.

Environ Microbiol 21:3539–3547.

7. Weiss MC, Sousa FL, Mrnjavac N, Neukirchen S, Roettger M, Nelson-Sathi S, Martin WF. 2016. The physiology and habitat of the last universal common ancestor. Nat Microbiol 1:1–8.
8. David LA, Alm EJ. 2011. Rapid evolutionary innovation during an Archaean genetic expansion. Nature 469:93–96.
9. Canfield DE, Rosing MT, Bjerrum C. 2006. Early anaerobic metabolisms. Philos Trans R Soc B Biol Sci 361:1819–1836.
10. Knoll AH, Bergmann KD, Strauss J V. 2016. Life: the first two billion years. Philos Trans R Soc B Biol Sci 371:20150493.
11. Bar-On YM, Phillips R, Milo R. 2018. The biomass distribution on Earth. Proc Natl Acad Sci U S A 115:6506–6511.
12. Stewart EJ. 2012. Growing unculturable bacteria. J Bacteriol 194:4151–4160.
13. Radajewski S, Ineson P, Parekh NR, Murrell JC. 2000. Stable-isotope probing as a tool in microbial ecology. Nature 403:646–649.
14. Solden L, Lloyd K, Wrighton K. 2016. The bright side of microbial dark matter: Lessons learned from the uncultivated majority. Curr Opin Microbiol 31:217–226.
15. Lloyd KG, Steen AD, Ladau J, Yin J, Crosby L. 2018. Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. mSystems 3:1–12.
16. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U, Brodie EL, Williams KH, Hubbard SS, Banfield JF. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical

processes in an aquifer system. Nat Commun 7:13219.

17. Baker BJ, Lazar CS, Teske AP, Dick GJ. 2015. Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. Microbiome 3:14.

18. Schimel J. 2016. Microbial ecology: Linking omics to biogeochemistry. Nat Microbiol 1:1–2.

19. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, d'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F, Lepoivre C, Lima-Mendez G, Poulain J, Poulos BT, Royo-Llonch M, Sarmento H, Vieira-Silva S, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Bowler C, de Vargas C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Not F, Ogata H, Pesant S, Speich S, Stemmann L, Sullivan MB, Weissenbach J, Wincker P, Karsenti E, Raes J, Acinas SG, Bork P, Boss E, Bowler C, Follows M, Karp-Boss L, Krzic U, Reynaud EG, Sardet C, Sieracki M, Velayoudon D. 2015. Structure and function of the global ocean microbiome. Science (80-) 348:1261359–1261359.

20. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. 2017. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 35:833–844.

21. Hardwick SA, Chen WY, Wong T, Kanakamedala BS, Deveson IW, Ongley SE, Santini NS, Marcellin E, Smith MA, Nielsen LK, Lovelock CE, Neilan BA, Mercer TR. 2018. Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis. Nat Commun 9:1–10.

22. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2019. GenBank. Nucleic Acids Res 47:D94–D99.

23. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–D745.

24. Mavromatis K, Ivanova NN, Chen I-MA, Szeto E, Markowitz VM, Kyrpides NC. 2009. The DOE-JGI Standard Operating Procedure for the Annotations of Microbial Genomes. SIGS 1:63–67.

25. Consortium TU. 2008. The Universal Protein Resource (UniProt). Nucleic Acids Res 36:D190–D195.

26. Royalty TM, Steen AD. 2019. Quantitatively Partitioning Microbial Genomic Traits among Taxonomic Ranks across the Microbial Tree of Life. mSphere 4:e00446-19.

27. Martiny JBH, Jones SE, Lennon JT, Martiny AC. 2015. Microbiomes in light of traits: A phylogenetic perspective. Science (80- ) 350:aac9323.

28. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A,

Hugenholtz P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 36:996–1004.

29. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol 2:1533–1542.

30. Royalty TM, Steen AD. 2019. Theoretical and Simulation-Based Investigation of the Relationship between Sequencing Effort, Microbial Community, Richness, and Diversity in Binning Metagenome-Assembled Genomes. mSystems 4:e00384-19.

31. Louca S, Martiny AC, Polz MF, Doolittle WF, Nelson MB, Hahn AS, Rubin I, Mazel F, Hawley AK, Connor MIO, Crowe SA, Hallam SJ. 2018. Function and functional redundancy in microbial communities. Nat Ecol Evol 2:936–943.

32. Middelburg JJ. 2019. Marine Carbon Biogeochemistry: A Primer for Earth System ScientistsSpringer Briefs in Earth System Sciences.

33. Houghtons RA. 2003. The Contemporary Carbon CycleBiogeochemistry: Treatise on Geochemistry, Volume 8.

34. Albert DB, Martens CS, Alperin MJ. 1998. Biogeochemical processes controlling methane in gassy coastal sediments - Part 2: Groundwater flow control of acoustic turbidity in Eckernforde Bay Sediments. Cont Shelf Res 18:1771–1793.

35. Martens CS, Albert DB, Alperin MJ. 1998. Biogeochemical processes controlling methane in gassy coastal sediments - Part 1. A model coupling organic matter flux to gas production, oxidation and transport. Cont Shelf Res 18:1741–1770.

36. Martiny AC, Treseder K, Pusch G. 2013. Phylogenetic conservatism of functional traits in

microorganisms. ISME J 7:830–838.

37. Zimmerman AE, Martiny AC, Allison SD. 2013. Microdiversity of extracellular enzyme genes among sequenced prokaryotic genomes. ISME J 7:1187–1199.

38. Federhen S. 2012. The NCBI Taxonomy database. Nucleic Acids Res 40:136–143.

39. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. Nat Microbiol 1:1–6.

40. Tatusov RL, Galperin MY, Natale DA, Koonin E V. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–36.

41. Philippot L, Andersson SGE, Battin TJ, Prosser JI, Schimel JP, Whitman WB, Hallin S. 2010. The ecological coherence of high bacterial taxonomic ranks. Nat Rev Microbiol 8:523–529.

42. Harrison XA, Donaldson L, Correa-Cano ME, Evans J, Fisher DN, Goodwin CED, Robinson BS, Hodgson DJ, Inger R. 2018. A brief introduction to mixed effects modelling and multi-model inference in ecology. PeerJ 6:e4794.

43. Danczak RE, Johnston MD, Kenah C, Slattery M, Wrighton KC, Wilkins MJ. 2017. Members of the Candidate Phyla Radiation are functionally differentiated by carbon- and nitrogen-cycling capabilities. Microbiome 5:112.

44. Emerson D, Fleming EJ, McBeth JM. 2010. Iron-Oxidizing Bacteria: An Environmental and Genomic Perspective. Annu Rev Microbiol 64:561–583.

45. Singh AH, Doerks T, Letunic I, Raes J, Bork P. 2009. Discovering functional novelty in

metagenomes: Examples from light-mediated processes. J Bacteriol 91:32–41.

46. Mendler K, Chen H, Parks DH, Lobb B, Hug LA, Doxey AC. 2019. AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. Nucleic Acids Res 4567:1–7.

47. Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: Building the web of life. Nat Rev Genet 16:472–482.

48. Schloter M, Lebuhn M, Heulin T, Hartmann A. 2000. Ecology and evolution of bacterial microdiversity. FEMS Microbiol Rev 24:647–660.

49. Moore EK, Jelen BI, Giovannelli D, Raanan H, Falkowski PG. 2017. Metal availability and the expanding network of microbial metabolisms in the Archaean eon. Nat Geosci 10:629–636.

50. Koeppel AF, Wu M. 2012. Lineage-dependent ecological coherence in bacteria. FEMS Microbiol Ecol 81:574–582.

51. Min CK, Yang JS, Kim S, Choi MS, Kim IS, Cho NH. 2008. Genome-based construction of the metabolic pathways of Orientia tsutsugamushi and comparative analysis within the Rickettsiales order. Comp Funct Genomics 2008.

52. Wolgemuth CW. 2015. Flagellar motility of the pathogenic spirochetes. Semin Cell Dev Biol 46:104–112.

53. Dreyer M, Aeby S, Oevermann A, Greub G. 2015. Prevalence and diversity of Chlamydiales in Swiss ruminant farms. Pathog Dis 73:1–4.

54. Wexler HM. 2007. Bacteroides: The good, the bad, and the nitty-gritty. Clin Microbiol Rev 20:593–621.

55. Moran NA. 2002. Microbial Minimalism: Genome Reduction in Bacterial Pathogens. Cell 108:583–586.

56. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119.

57. Marchler-Bauer A, Bryant SH. 2004. CD-Search: Protein domain annotations on the fly. Nucleic Acids Res 32:W327–W331.

58. Leimbach A, Poehlein A, Witten A, Wellnitz O, Shpigel N, Petzl W, Zerbe H, Daniel R, Dobrindt U. 2016. Whole-Genome Draft Sequences of Six Commensal Fecal and Six Mastitis-Associated *Escherichia coli* Strains of Bovine Origin: TABLE 1. Genome Announc 4:e00753-16.

59. Wood S. 2017. mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation.

60. Venables WN, Ripley BD. 2002. Modern Applied Statistics with S. 7.3.49. Springer, New York.

61. R Core Team. 2018. R: A Language and Environmen for Statistical Computing. Vienna, Austria.

62. Philip D. 2003. Computer program review VEGAN , a package of R functions for community ecology. J Veg Sci 14:927–930.

63. J P, Bates D, S D, D S, R Core Team. 2019. nlme: linear and nonlinear mixed effects models. R package version 3.1-141.

64. Knight R, Navas J, Quinn RA, Sanders JG, Zhu Q, Vrbanac A, Taylor BC, Aksenov A,

Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall L-I, McDonald D, Melnik A V, Morton JT, Navas J, Quinn RA, Sanders JG, Swafford AD, Thompson LR, Tripathi A, Xu ZZ, Zaneveld JR, Zhu Q, Caporaso JG, Dorrestein PC. 2018. Best practices for analysing microbiomes. Nat Rev Microbiol.

65. Tully BJ, Graham ED, Heidelberg JF. 2018. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data 5:1–8.

66. Lander ES, Waterman MS. 1988. Genomic mapping by end-characterized random clones: a mathematical analysis. Genomics 231–239.

67. Wendl MC, Kota K, Weinstock GM, Mitreva M. 2013. Coverage theories for metagenomic DNA sequencing based on a generalization of Stevens' theorem. J Math Biol 67:1141–1161.

68. Rodriguez-R LM, Konstantinidis KT. 2014. Estimating coverage in metagenomic data sets and why it matters. ISME J 8:2349–2351.

69. Pielou EC. 1966. The measurement of diversity in different types of biological collections. J Theor Biol 13:131–144.

70. Zwietering MH, Jongenburger I, Rombouts FM, Van't Riet K. 1990. Modeling of the bacterial growth curve. Appl Environ Microbiol 56:1875–1881.

71. Carini P, Marsden PJ, Leff JW, Morgan EE, Strickland MS, Fierer N. 2016. Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. Nat Microbiol 2:1–6.

72. Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW. 2015. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics 15:141–161.

73. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu WT, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, Lapidus A, Meyer F, Yilmaz P, Parks DH, Eren AM, Schriml L, Banfield JF, Hugenholtz P, Woyke T. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 35:725–731.

74. Karsenti E, Acinas SG, Bork P, Bowler C, de Vargas C, Raes J, Sullivan M, Arendt D, Benzoni F, Claverie JM, Follows M, Gorsky G, Hingamp P, Iudicone D, Jaillon O, Kandels-Lewis S, Krzic U, Not F, Ogata H, Pesant S, Reynaud EG, Sardet C, Sieracki ME, Speich S, Velayoudon D, Weissenbach J, Wincker P. 2011. A holistic approach to marine Eco-systems biology. PLoS Biol 9:7–11.

75. Howe AC, Jansson JK, Malfatti SA, Tringe SG, Tiedje JM, Brown CT. 2014. Tackling soil diversity with the assembly of large, complex metagenomes. Proc Natl Acad Sci 111:4904–4909.

76. Schirmer M, Smeekens SP, Vlamakis H, Jaeger M, Oosting M, Franzosa EA, Jansen T, Jacobs L, Bonder MJ, Kurilshikov A, Fu J, Joosten LAB, Zhernakova A, Huttenhower C, Wijmenga C, Netea MG, Xavier RJ. 2016. Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. Cell 167:1125-1136.e8.

77. Roesch LFW, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, Kent AD, Daroub SH, Camargo FAO, Farmerie WG, Triplett EW. 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J 1:283–290.

78. Feng BW, Li XR, Wang JH, Hu ZY, Meng H, Xiang LY, Quan ZX. 2009. Bacterial diversity of water and sediment in the Changjiang estuary and coastal area of the East China Sea. FEMS Microbiol Ecol 70:236–248.

79. Rintala A, Pietilä S, Munukka E, Eerola E, Pursiheimo JP, Laiho A, Pekkala S, Huovinen P. 2017. Gut microbiota analysis results are highly dependent on the 16s rRNA gene target region, whereas the impact of DNA extraction is minor. J Biomol Tech 28:19–30.

80. Kang S, Rodrigues JLM, Ng JP, Gentry TJ. 2016. Hill number as a bacterial diversity measure framework with high-throughput sequence data. Sci Rep 6:1–4.

81. Zaheer R, Noyes N, Ortega Polo R, Cook SR, Marinier E, Van Domselaar G, Belk KE, Morley PS, McAllister TA. 2018. Impact of sequencing depth on the characterization of the microbiome and resistome. Sci Rep 8:5890.

82. Ghurye JS, Cepeda-Espinoza V, Pop M. 2016. Metagenomic assembly: Overview, challenges and applications. Yale J Biol Med 89:353–362.

83. Lonardi S, Mirebrahim H, Wanamaker S, Alpert M, Ciardo G, Duma D, Close TJ. 2015. When less is more: “Slicing” sequencing data improves read decoding accuracy and de novo assembly quality. Bioinformatics 31:2972–2980.

84. Rosselló-Móra R, Amann R. 2015. Past and future species definitions for Bacteria and Archaea. Syst Appl Microbiol 38:209–216.

85. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP,

Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IMJ, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar S V., Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DMD, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Catenazzi MCE, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo K V., Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov D V., Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM,

Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59.

86. Flajolet P, Gardy D, Thimonier L. 1992. Birthday paradox, coupon collectors, caching algorithms and self-organizing search. Discret Appl Math 39:207–229.

87. Magurran AE. 1988. Ecological Diversity and Its Measurement, 1st ed. Croom Helm Ltd.

88. Galand PE, Casamayor EO, Kirchman DL, Lovejoy C. 2009. Ecology of the rare microbial biosphere of the Arctic Ocean. Proc Natl Acad Sci 106:22427–22432.

89. Shannon CE. 1948. A mathematical theory of communication. Bell Syst Tech J 27:379–423.

90. Leinonen R, Sugawara H, Shumway M. 2010. The Sequence Read Archive. Nucleic Acids Res 39:2010–2012.

91. Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863–864.

92. Graham ED, Heidelberg JF, Tully BJ. 2017. BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. PeerJ 5:e3035.

93. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina

sequence data. Bioinformatics 30:2114–20.

94. Li D, Luo R, Liu C-M, Ting H-F, Sadakane K, Yamashita H, Lam T-W. 2016. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods 102:3–11.

95. Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152.

96. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–9.

97. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–55.

98. Royalty TM, Steen AD. 2020. A quantitative measure of functional redundancy in microbial ecosystems. bioRxiv.

99. Coles VJ, Stukel MR, Brooks MT, Burd A, Crump BC, Moran MA, Paul JH, Satinsky BM, Yager PL, Zielinski BL, Hood RR. 2017. Ocean biogeochemistry modeled with emergent trait-based genomics. Science (80- ) 358:1149–1154.

100. Naeem S. 1998. Species redundancy and ecosystem reliability. Conserv Biol 12:39–45.

101. Hunting ER, Vijver MG, van der Geest HG, Mulder C, Kraak MHS, Breure AM, Admiraal W. 2015. Resource niche overlap promotes stability of bacterial community metabolism in experimental microcosms. Front Microbiol 6:1–7.

102. Teeling H, Fuchs BM, Becher D, Klockow C, Gardebrecht A, Bennke CM, Kassabgy M, Huang S, Mann AJ, Waldmann J, Weber M, Klindworth A, Otto A, Lange J, Bernhardt J,

Reinsch C, Hecker M, Peplies J, Bockelmann FD, Callies U, Gerdts G, Wichels A, Wiltshire KH, Glöckner FO, Schweder T, Amann R. 2012. Substrate-Controlled Succession of Marine Bacterioplankton Populations Induced by a Phytoplankton Bloom. Science (80- ) 336:608–611.

103. Heintz-Buschart A, Wilmes P. 2018. Human Gut Microbiome: Function Matters. Trends Microbiol 26:563–574.

104. Rao R. 1982. Diversity and Dissimilarity Coefficients: A Unified Approach. Theor Popul Biol 21:24–43.

105. Ricotta C, de Bello F, Moretti M, Caccianiga M, Cerabolini BEL, Pavoine S. 2016. Measuring the functional redundancy of biological communities: a quantitative guide. Methods Ecol Evol 7:1386–1395.

106. Mouillot D, Villéger S, Parravicini V, Kulbicki M, Arias-González JE, Bender M, Chabanet P, Floeter SR, Friedlander A, Vigliola L, Bellwood DR. 2014. Functional over-redundancy and high functional vulnerability in global fish faunas on tropical reefs. Proc Natl Acad Sci U S A 111:13757–13762.

107. Ricotta C, Szeidl L. 2006. Towards a unifying approach to diversity measures: Bridging the gap between the Shannon entropy and Rao's quadratic index. Theor Popul Biol 70:237–243.

108. Villéger S, Mason NWH, Mouillot D. 2008. New Mulitdimensional Functional Diversity Indicies for a Muiltifaceted Framework in Functional Ecology. Ecology 39:2290–2301.

109. Staley J, Konopka A. 1985. Measurement of In Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. Annu Rev Microbiol 39:321–346.

110. Allison SD, Martiny JBH. 2009. Resistance, resilience, and redundancy in microbial communities. Light Evol 2:149–166.

111. Delgado-Baquerizo M, Giaramida L, Reich PB, Khachane AN, Hamonts K, Edwards C, Lawton LA, Singh BK. 2016. Lack of functional redundancy in the relationship between microbial diversity and ecosystem functioning. J Ecol 104:936–946.

112. Krause S, Le Roux X, Niklaus PA, van Bodegom PM, Lennon T. JT, Bertilsson S, Grossart HP, Philippot L, Bodelier PLE. 2014. Trait-based approaches for understanding microbial biodiversity and ecosystem functioning. Front Microbiol 5:1–10.

113. Patil GP, Taillie C. 1982. Diversity as a concept and its measurement. J Am Stat Assoc 77:548–561.

114. McGill BJ, Etienne RS, Gray JS, Alonso D, Anderson MJ, Benecha HK, Dornelas M, Enquist BJ, Green JL, He F, Hurlbert AH, Magurran AE, Marquet PA, Maurer BA, Ostling A, Soykan CU, Ugland KI, White EP. 2007. Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework. Ecol Lett 10:995–1015.

115. Pajares S, Ramos R. 2019. Processes and Microorganisms Involved in the Marine Nitrogen Cycle: Knowledge and Gaps. Front Mar Sci 6.

116. Malik AA, Martiny JBH, Brodie EL, Martiny AC, Treseder KK, Allison SD. 2019. Defining trait-based microbial strategies with consequences for soil carbon cycling under climate change. ISME J 445866.

117. Hubert C. 2009. A Constant Flux of Diverse. Program 1541:1541–1544.

118. McPherson JM, Yeager LA, Baum JK. 2018. A simulation tool to scrutinise the behaviour

of functional diversity metrics. Methods Ecol Evol 9:200–206.

119. Tilman D. 2001. Functional Diversity. Encycl Biodivers Second Ed 3:587–596.

120. Wuchter C, Abbas B, Coolen MJL, Herfort L, Van Bleijswijk J, Timmers P, Strous M, Teira E, Herndl GJ, Middelburg JJ, Schouten S, Damsté JSS. 2006. Archaeal nitrification in the ocean. Proc Natl Acad Sci U S A 103:12317–12322.

121. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Xu ZZ, Jiang L, Haroon MF, Kanbar J, Zhu Q, Song SJ, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R, Agosto Rivera JL, Al-Moosawi L, Alverdy J, Amato KR, Andras J, Angenent LT, Antonopoulos DA, Apprill A, Armitage D, Ballantine K, Bárta J, Baum JK, Berry A, Bhatnagar A, Bhatnagar M, Biddle JF, Bittner L, Boldgiv B, Bottos E, Boyer DM, Braun J, Brazelton W, Brearley FQ, Campbell AH, Caporaso JG, Cardona C, Carroll JL, Cary SC, Casper BB, Charles TC, Chu H, Claar DC, Clark RG, Clayton JB, Clemente JC, Cochran A, Coleman ML, Collins G, Colwell RR, Contreras M, Crary BB, Creer S, Cristol DA, Crump BC, Cui D, Daly SE, Davalos L, Dawson RD, Defazio J, Delsuc F, Dionisi HM, Dominguez-Bello MG, Dowell R, Dubinsky EA, Dunn PO, Ercolini D, Espinoza RE, Ezenwa V, Fenner N, Findlay HS, Fleming ID, Fogliano V, Forsman A, Freeman C, Friedman ES, Galindo G, Garcia L, Garcia-Amado MA, Garshelis D, Gasser RB, Gerdts G, Gibson MK, Gifford I, Gill RT, Giray T, Gittel A, Golyshin P, Gong D,

Grossart HP, Guyton K, Haig SJ, Hale V, Hall RS, Hallam SJ, Handley KM, Hasan NA, Haydon SR, Hickman JE, Hidalgo G, Hofmockel KS, Hooker J, Hulth S, Hultman J, Hyde E, Ibáñez-Álamo JD, Jastrow JD, Jex AR, Johnson LS, Johnston ER, Joseph S, Jurburg SD, Jurelevicius D, Karlsson A, Karlsson R, Kauppinen S, Kellogg CTE, Kennedy SJ, Kerkhof LJ, King GM, Kling GW, Koehler A V., Krezalek M, Kueneman J, Lamendella R, Landon EM, Lanede Graaf K, LaRoche J, Larsen P, Laverock B, Lax S, Lentino M, Levin II, Liancourt P, Liang W, Linz AM, Lipson DA, Liu Y, Lladser ME, Lozada M, Spirito CM, MacCormack WP, MacRae-Crerar A, Magris M, Martín-Platero AM, Martín-Vivaldi M, Martínez LM, Martínez-Bueno M, Marzinelli EM, Mason OU, Mayer GD, McDevitt-Irwin JM, McDonald JE, McGuire KL, McMahon KD, McMinds R, Medina M, Mendelson JR, Metcalf JL, Meyer F, Michelangeli F, Miller K, Mills DA, Minich J, Mocali S, Moitinho-Silva L, Moore A, Morgan-Kiss RM, Munroe P, Myrold D, Neufeld JD, Ni Y, Nicol GW, Nielsen S, Nissimov JI, Niu K, Nolan MJ, Noyce K, O'Brien SL, Okamoto N, Orlando L, Castellano YO, Osuolale O, Oswald W, Parnell J, Peralta-Sánchez JM, Petraitis P, Pfister C, Pilon-Smits E, Piombino P, Pointing SB, Pollock FJ, Potter C, Prithiviraj B, Quince C, Rani A, Ranjan R, Rao S, Rees AP, Richardson M, Riebesell U, Robinson C, Rockne KJ, Rodriguezl SM, Rohwer F, Roundstone W, Safran RJ, Sangwan N, Sanz V, Schrenk M, Schrenzel MD, Scott NM, Seger RL, Seguinorlando A, Seldin L, Seyler LM, Shakhsheer B, Sheets GM, Shen C, Shi Y, Shin H, Shogan BD, Shutler D, Siegel J, Simmons S, Sjöling S, Smith DP, Soler JJ, Sperling M, Steinberg PD, Stephens B, Stevens MA, Taghavi S, Tai V, Tait K, Tan CL, Taş N, Taylor DL, Thomas T, Timling I, Turner BL, Urich T, Ursell LK, Van Der Lelie D, Van Treuren W, Van

Zwieten L, Vargas-Robles D, Thurber RV, Vitaglione P, Walker DA, Walters WA, Wang S, Wang T, Weaver T, Webster NS, Wehrle B, Weisenhorn P, Weiss S, Werner JJ, West K, Whitehead A, Whitehead SR, Whittingham LA, Willerslev E, Williams AE, Wood SA, Woodhams DC, Yang Y, Zaneveld J, Zarraonaindia I, Zhang Q, Zhao H. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457–463.

122. Steen AD, Crits-Christoph A, Carini P, DeAngelis KM, Fierer N, Lloyd KG, Cameron Thrash J. 2019. High proportions of bacteria and archaea across most biomes remain uncultured. ISME J 13:3126–3130.

123. Musat N, Foster R, Vagner T, Adam B, Kuypers MMM. 2012. Detecting metabolic activities in single cells, with emphasis on nanoSIMS. FEMS Microbiol Rev 36:486–511.

124. Bairoch A, Apweiler R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28:45–48.

125. Froelich PN, Klinkhammer GP, Bender ML, Luedtke NA, Heath GR, Cullen D, Dauphin P, Hammond D, Hartman B, Maynard V. 1979. Early oxidation of organic matter in pelagic sediments of the eastern equatorial Atlantic: suboxic diagenesis. Geochim Cosmochim Acta 43:1075–1090.

126. LaRowe DE, Arndt S, Bradley JA, Estes ER, Hoarfrost A, Lang SQ, Lloyd KG, Mahmoudi N, Orsi WD, Shah Walter SR, Steen AD, Zhao R. 2020. The fate of organic carbon in marine sediments - New insights from recent data and analysis. Earth-Science Rev 204:103146.

127. Hutchings JA, Bianchi TS, Najjar RG, Herrmann M, Kemp WM, Hinson AL, Feagin RA. 2020. Carbon Deposition and Burial in Estuarine Sediments of the Contiguous United

States. Global Biogeochem Cycles 34.

128. Zepp RG, Schlotzhauer PF. 1981. Comparison of photochemical behavior of various humic substances in water: III. Spectroscopic properties of humic substances. Chemosphere 10:479–486.

129. Martens S, Goldhaber B. 1978. Early diagenesis in transitional sedimentary environments of the White Oak River Estuary , North Carolina 23.

130. Kelley C a., Martens CS, Chanton JP. 1990. Variations in sedimentary carbon remineralization rates in the White Oak River estuary, North Carolina. Limnol Oceanogr 35:372–383.

131. Lloyd KG, Alperin MJ, Teske A. 2011. Environmental evidence for net methane production and oxidation in putative ANaerobic MEthanotrophic (ANME) archaea. Environ Microbiol 13:2548–2564.

132. Steen AD, Kevorkian RT, Bird JT, Dombrowski N, Baker BJ, Hagen SM, Mulligan KH, Schmidt JM, Webber AT, Royalty TM, Alperin MJ. 2019. Kinetics and identities of extracellular peptidases in subsurface sediments of the White Oak River Estuary, NC. Appl Environ Microbiol.

133. Kogan I, Paull CK. 2005. Coastal seismic wipe-outs: Distribution controlled by pore water. Mar Geol 217:161–175.

134. Lazar CS, Baker BJ, Seitz KW, Teske AP. 2017. Genomic reconstruction of multiple lineages of uncultured benthic archaea suggests distinct biogeochemical roles and ecological niches. ISME J 11:1118–1129.

135. Okie JG, Horn DJ Van, Storch D, Barrett JE, Gooseff MN, Kopsova L, Takacs-vesbach

CD, Okie JG. 2015. Niche and metabolic principles explain patterns of diversity and distribution : theory and a case study with soil bacterial communities. R Soc Proc B.

136. Morrissey EM, Mau RL, Hayer M, Liu XJA, Schwartz E, Dijkstra P, Koch BJ, Allen K, Blazewicz SJ, Hofmockel K, Pett-Ridge J, Hungate BA. 2019. Evolutionary history constrains microbial traits across environmental variation. Nat Ecol Evol 3:1064–1069.

137. Knab NJ, Dale AW, Lettmann K, Fossing H, Jørgensen BB. 2008. Thermodynamic and kinetic control on anaerobic oxidation of methane in marine sediments. Geochim Cosmochim Acta 72:3746–3757.

138. Land LA, Paull CK. 2001. Thermal gradients as a tool for estimating groundwater advective rates in a coastal estuary: White Oak River, North Carolina, USA. J Hydrol 248:198–215.

139. Winner MD, Coble RW. 1996. Hydrogeologic framework of the North Carolina Coastal Plain. US Geol Surv Prof Pap 11–150.

140. Yamamoto S, Alcauskas JB, Crozier TE. 1976. Solubility of Methane in Distilled Water and Seawater. J Chem Eng Data 21:78–80.

141. Helms JR, Stubbins A, Ritchie JD, Minor EC, Kieber DJ, Mopper K. 2009. Erratum: Absorption spectral slopes and slope ratios as indicators of molecular weight, source, and photobleaching of chromophoric dissolved organic matter. Limnol Oceanogr 54:1023.

142. Kelley CA, Martens CS, Chanton JP. 1990. Variations in sedimentary carbon remineralization rates in the White Oak River estuary, North Carolina. Limnol Oceanogr 35:372–383.

143. Boudreau BP. 1997. Diagenetic Models and Their Implementation. Springer.

144. Oelkers EH. 1991. Calculation of diffusion coefficients for aqueous organic species at temperatures from 0 to 350 °C. Geochim Cosmochim Acta 55:3515–3529.

145. Lapham LL, Alperin M, Chanton J, Martens C. 2008. Upward advection rates and methane fluxes, oxidation, and sources at two Gulf of Mexico brine seeps. Mar Chem 112:65–71.

146. Hartnett HE, Keil RG, Hedges JI, Devol AH. 1998. Influence of oxygen exposure time on organic carbon preservation in continental margin sediments. Nature 391:572–574.

147. Middelburg JJ, Vlug T, Jaco F, van der Nat WA. 1993. Organic matter mineralization in marine systems. Glob Planet Change 8:47–58.

148. Bianchi TS, Canuel EA. 2011. Chemical biomarkers in aquatic ecosystemsChemical Biomarkers in Aquatic Ecosystems.

149. Megonigal JP, Hines ME, Visscher PT. 2003. Anaerobic Metabolism: Linkages to Trace Gases and Aerobic ProcessesBiogeochemistry: Treatise on Geochemistry, Volume 8.

150. Jørgensen BB, Findlay AJ, Pellerin A. 2019. The biogeochemical sulfur cycle of marine sediments. Front Microbiol 10:1–27.

# VITA

Taylor M. Royalty was born in Raleigh, N.C., USA. He obtained his B.S. in Biology at Western Carolina University, Cullowhee, N.C., USA. While at Western Carolina University, Taylor studied aqueous copper-sulfide chemistry under the advisement of Dr. Jack Summers. Taylor proceeded to obtain his M.S. in Atmospheric Science at North Carolina State University, Raleigh, N.C., USA. While at North Carolina State University, Taylor studied marine aerosol hygroscopicity and atmospheric aerosol fluxes under the co-advisement of Dr. Markus Petters and Dr. Nicholas Meskhidze. Taylor pursued his Ph.D. at the University of Tennessee, Knoxville, USA. While at the University of Tennessee, Knoxville, Taylor studied diversity in microbial communities, taxonomic groups, and community function under the advisement of Dr. Andrew Steen.