A MULTI-OMICS APPROACH TO ASSESSING GROWTH, STRESS, AND DISTURBANCE IN SOIL MICROBIAL COMMUNITIES

By Peter Francis Chuckran

A Dissertation

Submitted for the Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

in Biology

Northern Arizona University

May 2022

Approved:

Paul Dijkstra, Ph.D., Chair

Bruce A Hungate, Ph.D.

Egbert Schwartz, Ph.D.

Steven J. Blazewicz, Ph.D.

ABSTRACT

A MULTI-OMICS APPROACH TO ASSESSING GROWTH, STRESS, AND DISTURBANCE IN SOIL MICROBIAL COMMUNITIES

PETER FRANCIS CHUCKRAN

Microbes in soil are well-known drivers of several ecosystem processes, yet our ability to study their genetic controls on a community level is relatively recent. The total DNA and RNA of a microbial community—referred to as the metagenome and metatranscriptome and collectively part of the field of study known as "omics"—can yield valuable insight into microbial physiology and function, community structure, and the evolutionary processes of microorganisms. This dissertation leverages metagenomics and metatranscriptomics to assess soil microbial communities with a particular focus on understanding how this approach can be used to better understand dimensions of growth, stress, and disturbance. The first chapter introduces this topic and reviews the current state of the literature and crucial knowledge gaps, as well as a brief description of the subsequent chapters. Chapter 2 describes an experiment where we observed the transcriptional controls of soil microbial communities in response to labile carbon inputs and found that inputs of glucose rapidly stimulated the transcription of nitrogen cycling genes. Chapter 3 is a broad-scale data analysis of genomic traits in bacterial communities from soil, marine, host-associated, and hot-spring microbial communities. We found that soil communities have relationships between genomic traits which are distinct from those in other ecosystems—indicating a unique set of selection pressures in soils. In Chapter 4 we follow-up on these results and examine the distribution of genomic traits in soils along multiple environmental parameters. This analysis showed that bacterial traits in soils are likely driven by carbon

limitation and soil pH. In Chapter 5 we reexamine the transcriptional response described in Chapter 2, this time focusing on how genomic traits such as nucleotide and codon selection impact the short-term response of soil microbes during growth and stress. Together, these results highlight the numerous ways in which we can derive insights from multi-omics data and how these findings can enhance our understanding of microbial life in soils.

ACKNOWLEDGEMENTS

This work would not be possible without the support from the following people. I would like to thank Paul Dijkstra for his time, dedication, and patience as an advisor. Bruce Hungate for his valuable input and guidance as a co-advisor. Egbert Schwartz, and Steven Blazewicz for their feedback and intellectual contributions. Ember Morissey and Jeth Walkup for their assistance in sample collection and writing. Viacheslave Fofanov and Ella Sieradzki for their contributions to the bioinformatics analysis. Cody Flagg, Austin Rutherford, Jefferey Propster, and Jennifer Pett-Ridge for their assistance in the analysis and writing of Chapter 4. The LIMES lab for their valuable feedback: Ayla Martinez. Rachel Rubin, Bri Finley, Alicia Purcell, Megan Foley, Michaela Hayer, Rebecca Mau, Victoria Monsaint-Queeney, Raina Fitzpatrick, and Ben Koch. My friends and colleagues Anita Antoninka, Lydia Bailey, Henry Grover, Hannah Lee, Julia Stuart, Andrew Sanchez, Jesse Duff-Woodruff, Maggie Dogowitz, and Carl Roybal for their support and intellection contributions. Alan and Christina Chuckran, and Nara Bopp for their continued support.

TABLE OF CONTENTS

LIST OF FIGURES

DEDICATION

This dissertation is dedicated to Dr. Chester Orban, Dr. Douglas Frank, and Dr. Anita Antoninka.

PREFACE

These chapters are formatted as manuscripts to be submitted for publication in academic journals. Chapter 1 is a general introduction and review of the current literature and methodological approaches with citations formatted according to the Chicago Manual of Style. Chapter 2 is formatted for *mSystems*, where it was published in 2021. Chapter 3 is formatted for *FEMS Microbes* where it was published in 2022. Chapter 4 is formatted as a "Brief Communication" article for *ISMEJ* and is also available as a preprint via *bioRxiv*. Chapter 5 is formatted as a general research article with Chicago Manual of Style citation format.

CHAPTERS

CHAPTER 1
INTRODUCTION

Microbes in the soil dictate nutrient cycling (Wagg et al. 2019), sequester carbon from the atmosphere (Crowther et al. 2019), form crucial relationships with plants (Compant et al. 2019), and drive soil formation (Rillig and Mummey 2006). Although the importance of soil microorganisms to ecosystem function and global ecological processes has long been recognized, our ability to explore their genetic controls *en masse* is relatively recent. The cost of sequencing has dropped precipitously over the past decade and will likely only continue to become cheaper and more accessible (Tedersoo et al. 2021). This has resulted in the mass generation of metagenomic and metatranscriptomic data, effectively creating endless possibilities for analysis and a new computational challenge. Uncovering useful, concise, and informative metrics for assessing microbial communities is a central challenge in microbial ecology (Fierer, Wood, and Bueno de Mesquita 2021) and determining informative metrics for the assessment of "omics" data is similarly of high importance .

This dissertation uses a number of approaches to probe multi-omics datasets with a particular focus on how these approaches may be used to identify trends in growth, stress, and disturbance. Growth, stress, and disturbance are factors crucial to microbial functionality and are consequently the focus of countless studies in soil microbial ecology. Growth is an important and well-known dimension for understanding carbon use efficiency (Zheng et al. 2019; Dijkstra et al. 2015) and the formation/fate of soil organic carbon (Prommer et al. 2020; Hagerty et al. 2014), and is an important determinant for community dynamics (Hungate et al. 2021; Morrissey et al. 2016; Koch et al. 2018). Short-term disturbances, such as extreme temperatures, and long term stresses, such as drought or nutrient limitation, exert unique pressures on microbial communities

(Schimel, Balser, and Wallenstein 2007); and the resistance and resilience to disturbance and stress greatly influence community composition (Shade et al. 2012; Allison and Martiny 2008) and functionality (De Vries et al. 2012).

These themes were also chosen as they are at the heart of many of conceptual frameworks both in microbial and ecosystem-level ecology. Grime's 1977 Competitors-Stress Tolerators-Ruderal (C-S-R) framework, Malik et al. 2020 Yield-Acquisition-Stress (YAS) framework, copiotroph vs oligotroph (Lauro et al. 2009), and the r vs K strategist framework all rely, either partially or completely, on the response of an organism during growth, stress, and disturbance. Identifying metrics associated with these themes would therefore be an asset in determining how soil microbes fit into these conceptual frameworks and contribute to ecosystem function.

Throughout these chapters there will be occasional mixed-use of the term stress, in that it will sometimes refer to both stress and disturbance. This is a side-effect of microbial ecology being at the intersection of microbiology and ecosystem-level ecology. In ecology, a disturbance is more easily defined as an event which results in the partial mortality of an organism, such as herbivore grazing, whereas stress is more of a continuous condition which limits activity, such as drought (Grime 1977). However, bacteria are rarely subject to disturbance under this definition since it is hard to remove part of a unicellular organism without it's complete destruction (Schimel, Balser, and Wallenstein 2007). Further, the field of microbiology does not commonly use the term disturbance. Short-term changes in the environment which could be considered disturbances in macroecology would elicit a "stress-response" under microbial definition. Very likely the two fields never conferred with each other on these definitions, perhaps because they never imagined we would ever have such a great capacity to sequence environmental microbial

communities. Plante 2017 suggests the term *disturbance* refer to a short-term change in the environment which immediately impacts fitness, and *stress* refer to a more continuous limitation on growth. The introduction and discussion will use these definitions of stress and disturbance; however, the dissertation chapters, having been written for publication in more microbiology-focused journals (as opposed to ecology-focused journals), will often use "stress" in place of disturbance. The discussion will then put these results in context using Plante's definitions.

Description of Chapters

Short-term transcriptional responses in soil microbial communities are rarely observed and, consequently, much of our understanding of this process has been derived from bacteria in culture. Chapter 2 uses metatranscriptomics to determine the short-term response of microbial communities to additions of labile carbon. Since accessible carbon tends to be a limiting factor for the growth of soil heterotrophic bacteria (Hobbie and Hobbie 2013; Demoling, Figueroa, and Bååth 2007), a sudden input of simple sugars can rapidly stimulate microbial activity. The alleviation of carbon limitation results in nitrogen becoming the predominantly limiting nutrient and the rapid uptake of available nitrogen (Kamble and Bååth 2014). This is an important phenomenon as soil microorganisms are often subject to short-term pules of labile carbon, such as through litter leachate or root exudates (Kuzyakov and Blagodatskaya 2015). Chapter 2 aims to observe the transcriptional controls of this response, capturing both a dimension of growth (due to the stimulation of activity with labile carbon) and disturbance (with the sudden activation of nutrient limitation).

Similar to transcription, much of what we know about the genomic traits of soil microbes is based on findings from culture or marine systems. Genomic traits—such as genome size, GC

content, number of rRNA gene copies, codon usage, and number of regulatory genes—are potentially highly informatic metrics for assessing microbial life-strategy and function (Li et al. 2019; Roller, Stoddard, and Schmidt 2016); however, the distribution of traits between soil microbial communities has yet to be properly assessed. The study of genomic traits across systems could reveal important selection mechanisms as well as assist in informing a trait-based framework in microbial ecology. There has been increased interest in a trait-based framework for soil microbes (Westoby et al. 2021) and identifying easily accessible genomic traits would provide a valuable contribution towards this goal. Chapter 3 and 4 seek to close this gap with a series of large-scale metagenomic data analyses.

In Chapter 3, we assess the distribution of genomic traits in soil communities against those in marine, host-associated, and hot-spring environments. These systems were chosen as a point of comparison as they all have predominant drivers of genomic traits which might influence their distribution in predictable ways (Sabath et al. 2013; Batut et al. 2014; Giovannoni et al. 2005). For example, in marine systems nutrient limitation will often select for bacteria with reduced genomes and low GC content in order to curb the cost of reproduction (Giovannoni, Cameron Thrash, and Temperton 2014). Using community-derived averages for each trait, we compared over 100 metagenomes from the Joint Genome Institute (JGI) with the aims of: (1) determining if known relationships between genomic trait could be detected between communities (such as a positive relationship between genome size and GC content for marine communities) and (2) comparing these distributions to those of soils.

This idea is explored further in Chapter 4. There, we turn to a larger collection of soil metagenomes accessed from the National Ecological Observation Network (NEON). By assessing the relationship between genomic traits and environmental properties, we aimed to

uncover the predominant drivers of these traits in soil bacteria and the key mechanisms shaping these relationships. Specifically, we test whether nutrient stress, in the form of carbon limitation, could be driving genomic traits of soil bacterial communities.

Chapter 5 applies the concepts of the trait-based approach described in Chapters 3 and 4 to the metatranscriptomes in Chapter 2. Traits such as codon usage and GC content (Chen et al. 2016; Zhou et al. 2016) are known to influence rates of gene transcription. We use the metatranscriptome data described in Chapter 2 to see if these factors play an important role in growth and stress responses on the community-level. We also test whether these traits influence the transcription of bacterial genes in response to a sudden heat-shock. Through this analysis we aim to identify factors which contribute to growth and the response to disturbance.

These chapters describe several different analytical approaches for interpreting multi-omics data with a focus on identifying factors associated with specific life strategies and response mechanisms. It is my hope that by assessing themes of growth, stress, and disturbance with these approaches, this dissertation might reveal unique insight into the factors which dictate microbial life in soil and contribute to understanding how soil microbes contribute to ecosystem processes.

REFERENCES

Allison, Steven D., and Jennifer B.H. Martiny. 2008. “Resistance, Resilience, and Redundancy in Microbial Communities.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (SUPPL. 1): 11512–19. https://doi.org/10.1073/pnas.0801925105.

Batut, Bérénice, Carole Knibbe, Gabriel Marais, and Vincent Daubin. 2014. “Reductive Genome Evolution at Both Ends of the Bacterial Population Size Spectrum.” *Nature Reviews Microbiology* 12 (12): 841–50. https://doi.org/10.1038/nrmicro3331.

Chen, Wei Hua, Guanting Lu, Peer Bork, Songnian Hu, and Martin J. Lercher. 2016. “Energy Efficiency Trade-Offs Drive Nucleotide Usage in Transcribed Regions.” *Nature Communications* 7 (1): 1–10. https://doi.org/10.1038/ncomms11334.

Compant, Stéphane, Abdul Samad, Hanna Faist, and Angela Sessitsch. 2019. “A Review on the Plant Microbiome: Ecology, Functions, and Emerging Trends in Microbial Application.” *Journal of Advanced Research* 19 (September): 29–37. https://doi.org/10.1016/J.JARE.2019.03.004.

Crowther, T. W., J. van den Hoogen, J. Wan, M. A. Mayes, A. D. Keiser, L. Mo, C. Averill, and D. S. Maynard. 2019. “The Global Soil Community and Its Influence on Biogeochemistry.” *Science* 365 (6455). https://doi.org/10.1126/SCIENCE.AAV0550/ASSET/BC760EFF-B647-4105-B7F5-0A94D56382E8/ASSETS/GRAPHIC/365_AAV0550_F5.JPEG.

Demoling, Fredrik, Daniela Figueroa, and Erland Bååth. 2007. “Comparison of Factors Limiting Bacterial Growth in Different Soils.” *Soil Biology and Biochemistry* 39 (10): 2485–95. https://doi.org/10.1016/J.SOILBIO.2007.05.002.

Dijkstra, Paul, Elena Salpas, Dawson Fairbanks, Erin B. Miller, Shannon B. Hagerty, Kees Jan van Groenigen, Bruce A. Hungate, Jane C. Marks, George W. Koch, and Egbert Schwartz. 2015. “High Carbon Use Efficiency in Soil Microbial Communities Is Related to Balanced Growth, Not Storage Compound Synthesis.” *Soil Biology and Biochemistry* 89 (October): 35–43. https://doi.org/10.1016/j.soilbio.2015.06.021.

Fierer, Noah. 2017. “Embracing the Unknown: Disentangling the Complexities of the Soil Microbiome.” *Nature Reviews Microbiology* 15 (10): 579–90. https://doi.org/10.1038/nrmicro.2017.87.

Fierer, Noah, Stephen A. Wood, and Clifton P. Bueno de Mesquita. 2021. “How Microbes Can, and Cannot, Be Used to Assess Soil Health.” *Soil Biology and Biochemistry* 153 (February): 108111. https://doi.org/10.1016/J.SOILBIO.2020.108111.

Giovannoni, Stephen J., H. James Tripp, Scott Givan, Mircea Podar, Kevin L. Vergin, Damon Baptista, Lisa Bibbs, et al. 2005. “Genetics: Genome Streamlining in a Cosmopolitan Oceanic Bacterium.” *Science* 309 (5738): 1242–45. https://doi.org/10.1126/science.1114057.

Giovannoni, Stephen J, J Cameron Thrash, and Ben Temperton. 2014. “Implications of Streamlining Theory for Microbial Ecology.” *The ISME Journal* 8 (8): 1553–65. https://doi.org/10.1038/ismej.2014.60.

Grime, J P. 1977. “Evidence for the Existence of Three Primary Strategies in Plants and Its Relevance to Ecological and Evolutionary Theory.” *The American Naturalist* 111 (982): 1169–94. https://doi.org/10.1086/283244.

Hagerty, Shannon B., Kees Jan Van Groenigen, Steven D. Allison, Bruce A. Hungate, Egbert Schwartz, George W. Koch, Randall K. Kolka, and Paul Dijkstra. 2014. “Accelerated Microbial Turnover but Constant Growth Efficiency with Warming in Soil.” *Nature Climate Change 2014 4:10* 4 (10): 903–6. https://doi.org/10.1038/nclimate2361.

Hobbie, John E., and Erik A. Hobbie. 2013. “Microbes in Nature Are Limited by Carbon and Energy: The Starving-Survival Lifestyle in Soil and Consequences for Estimating Microbial Rates.” *Frontiers in Microbiology* 4 (November): 324. https://doi.org/10.3389/fmicb.2013.00324.

Hungate, Bruce A., Jane C. Marks, Mary E. Power, Egbert Schwartz, Kees Jan van Groenigen, Steven J. Blazewicz, Peter Chuckran, et al. 2021. “The Functional Significance of Bacterial Predators.” *MBio* 12 (2). https://doi.org/10.1128/MBIO.00466-21/SUPPL_FILE/MBIO.00466-21-SF001.PDF.

Kamble, Pramod N., and Erland Bååth. 2014. “Induced N-Limitation of Bacterial Growth in Soil: Effect of Carbon Loading and N Status in Soil.” *Soil Biology and Biochemistry* 74 (July): 11–20. https://doi.org/10.1016/J.SOILBIO.2014.02.015.

Koch, Benjamin J., Theresa A. McHugh, Michaela Hayer, Egbert Schwartz, Steven J. Blazewicz, Paul Dijkstra, Natasja van Gestel, et al. 2018. “Estimating Taxon-Specific Population Dynamics in Diverse Microbial Communities.” *Ecosphere* 9 (1): e02090. https://doi.org/10.1002/ecs2.2090.

Kuzyakov, Yakov, and Evgenia Blagodatskaya. 2015. “Microbial Hotspots and Hot Moments in Soil: Concept & Review.” *Soil Biology and Biochemistry* 83 (April): 184–99. https://doi.org/10.1016/J.SOILBIO.2015.01.025.

Lauro, Federico M, Diane McDougald, Torsten Thomas, Timothy J Williams, Suhelen Egan, Scott Rice, Matthew Z DeMaere, et al. 2009. “The Genomic Basis of Trophic Strategy in Marine Bacteria.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (37): 15527–33. https://doi.org/10.1073/pnas.0903507106.

Li, Junhui, Rebecca L. Mau, Paul Dijkstra, Benjamin J. Koch, Egbert Schwartz, Xiao-Jun Allen Liu, Ember M. Morrissey, et al. 2019. “Predictive Genomic Traits for Bacterial Growth in Culture versus Actual Growth in Soil.” *The ISME Journal*, May, 1. https://doi.org/10.1038/s41396-019-0422-z.

Malik, Ashish A., Jennifer B.H. Martiny, Eoin L. Brodie, Adam C. Martiny, Kathleen K. Treseder, and Steven D. Allison. 2020. "Defining Trait-Based Microbial Strategies with Consequences for Soil Carbon Cycling under Climate Change." *ISME Journal* 14 (1): 1–9. https://doi.org/10.1038/s41396-019-0510-0.

Morrissey, Ember M, Rebecca L Mau, Egbert Schwartz, J Gregory Caporaso, Paul Dijkstra, Natasja Van Gestel, Benjamin J Koch, et al. 2016. "Phylogenetic Organization of Bacterial Activity." *The ISME Journal* 1028: 2336–40. https://doi.org/10.1038/ismej.2016.28.

Plante, Craig J. 2017. "Defining Disturbance for Microbial Ecology." *Microbial Ecology* 74 (2): 259–63. https://doi.org/10.1007/s00248-017-0956-4.

Prommer, Judith, Tom W.N. Walker, Wolfgang Wanek, Judith Braun, David Zezula, Yuntao Hu, Florian Hofhansl, and Andreas Richter. 2020. "Increased Microbial Growth, Biomass, and Turnover Drive Soil Organic Carbon Accumulation at Higher Plant Diversity." *Global Change Biology* 26 (2): 669–81. https://doi.org/10.1111/GCB.14777.

Rillig, Matthias C., and Daniel L. Mummey. 2006. "Mycorrhizas and Soil Structure." *New Phytologist* 171 (1): 41–53. https://doi.org/10.1111/J.1469-8137.2006.01750.X.

Roller, Benjamin R.K., Steven F. Stoddard, and Thomas M. Schmidt. 2016. "Exploiting RRNA Operon Copy Number to Investigate Bacterial Reproductive Strategies." *Nature Microbiology* 1: 1–8. https://doi.org/10.1038/nmicrobiol.2016.160.

Sabath, Niv, Evandro Ferrada, Aditya Barve, and Andreas Wagner. 2013. "Growth Temperature and Genome Size in Bacteria Are Negatively Correlated, Suggesting Genomic Streamlining during Thermal Adaptation." *Genome Biology and Evolution* 5 (5): 966–77. https://doi.org/10.1093/gbe/evt050.

Schimel, Joshua, Teri C. Balser, and Matthew Wallenstein. 2007. "Microbial Stress-Response Physiology and Its Implications for Ecosystem Function." *Ecology* 88 (6): 1386–94. https://doi.org/10.1890/06-0219.

Shade, Ashley, Hannes Peter, Steven D. Allison, Didier L. Baho, Mercè Berga, Helmut Bürgmann, David H. Huber, et al. 2012. "Fundamentals of Microbial Community Resistance and Resilience." *Frontiers in Microbiology*. Frontiers Research Foundation. https://doi.org/10.3389/fmicb.2012.00417.

Tedersoo, Leho, Mads Albertsen, Sten Anslan, and Benjamin Callahan. 2021. "Perspectives and Benefits of High-Throughput Long-Read Sequencing in Microbial Ecology." *Applied and Environmental Microbiology* 87 (17): 1–19. https://doi.org/10.1128/AEM.00626-21/SUPPL_FILE/AEM.00626-21-S0001.PDF.

Vries, Franciska T. De, Mira E. Liiri, Lisa Bjørnlund, Matthew A. Bowker, Søren Christensen, Heikki M. Setälä, and Richard D. Bardgett. 2012. "Land Use Alters the Resistance and

Resilience of Soil Food Webs to Drought." *Nature Climate Change 2012 2:4* 2 (4): 276–80. https://doi.org/10.1038/nclimate1368.

Wagg, Cameron, Klaus Schlaeppi, Samiran Banerjee, Eiko E. Kuramae, and Marcel G.A. van der Heijden. 2019. "Fungal-Bacterial Diversity and Microbiome Complexity Predict Ecosystem Functioning." *Nature Communications 2019 10:1* 10 (1): 1–10. https://doi.org/10.1038/s41467-019-12798-y.

Westoby, Mark, Michael R. Gillings, Joshua S. Madin, Daniel A. Nielsen, Ian T. Paulsen, and Sasha G. Tetu. 2021. "Trait Dimensions in Bacteria and Archaea Compared to Vascular Plants." *Ecology Letters* 24 (7): 1487–1504. https://doi.org/10.1111/ele.13742.

Zheng, Qing, Yuntao Hu, Shasha Zhang, Lisa Noll, Theresa Böckle, Andreas Richter, and Wolfgang Wanek. 2019. "Growth Explains Microbial Carbon Use Efficiency across Soils Differing in Land Use and Geology." *Soil Biology and Biochemistry* 128 (January): 45–55. https://doi.org/10.1016/J.SOILBIO.2018.10.006.

Zhou, Zhipeng, Yunkun Danga, Mian Zhou, Lin Li, Chien Hung Yu, Jingjing Fu, She Chen, and Yi Liu. 2016. "Codon Usage Is an Important Determinant of Gene Expression Levels Largely through Its Effects on Transcription." *Proceedings of the National Academy of Sciences of the United States of America* 113 (41): E6117–25. https://doi.org/10.1073/pnas.1606724113.

CHAPTER 2

RAPID RESPONSE OF NITROGEN CYCLING GENE TRANSCRIPTION TO LABILE CARBON AMENDMENTS IN A SOIL MICROBIAL COMMUNITY

RUNNING TITLE: Rapid transcription of nitrogen cycling genes in soil

Authors:

Peter F. Chuckran[a]#, Viacheslav Fofanov[b,c], Bruce A. Hungate[a], Ember M Morrissey[d], Egbert Schwartz[a], Jeth Walkup[d], Paul Dijkstra[a]

[a] Center for Ecosystem Science and Society (ECOSS) and Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ, USA

[b] Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, USA

[c] School of Informatics, Computing and Cyber Systems, Northern Arizona University, Flagstaff, AZ, USA

[d] Division of Plant and Soil Sciences, West Virginia University, Morgantown, WV, USA

#Corresponding author: pfc25@nau.edu, pfchuckran@gmail.com

ABSTRACT

Episodic inputs of labile carbon (C) to soil can rapidly stimulate nitrogen (N) immobilization by soil microorganisms. However, the transcriptional patterns that underlie this process remain unclear. In order to better understand the regulation of N cycling in soil microbial communities, we conducted a 48 h laboratory incubation with an agricultural soil where we stimulated the uptake of inorganic N by amending the soil with glucose. We analyzed the metagenome and metatranscriptome of the microbial communities at four timepoints that corresponded with changes in N availability. The relative abundances of genes remained largely unchanged throughout the incubation. In contrast, glucose addition rapidly increased transcription of genes encoding for ammonium and nitrate transporters, enzymes responsible for N assimilation into biomass, and genes associated with the N regulatory network. This upregulation coincided with an increase in transcripts associated with glucose breakdown and oxoglutarate production, demonstrating a connection between C and N metabolism. When concentrations of ammonium were low, we observed a transient upregulation of genes associated with the nitrogen fixing enzyme nitrogenase. Transcripts for nitrification and denitrification were downregulated throughout the incubation, suggesting that dissimilatory transformations of N may be suppressed in response to labile C inputs in these soils. These results demonstrate that soil microbial communities can respond rapidly to changes in C availability by drastically altering the transcription of N cycling genes.

IMPORTANCE:

A large portion of activity in soil microbial communities occurs in short time frames in response to an increase in C availability, affecting the biogeochemical cycling of nitrogen. These changes

are of particular importance as nitrogen represents both a limiting nutrient for terrestrial plants as well as a potential pollutant. However, we lack a full understanding of the short-term effects of labile carbon inputs on the metabolism of microbes living in soil. Here, we found that soil microbial communities responded to labile carbon addition by rapidly transcribing genes encoding proteins and enzymes responsible for inorganic nitrogen acquisition, including nitrogen fixation. This work demonstrates that soil microbial communities respond within hours to carbon inputs through altered gene expression. These insights are essential for improved understanding of the microbial processes governing soil organic matter production, decomposition, and nutrient cycling in natural and agricultural ecosystems.

INTRODUCTION

Inorganic nitrogen (N) availability in soil dictates several ecosystem-level processes such as plant growth (1), greenhouse gas emissions in the form of nitrous oxide (2), and eutrophication from runoff (3). The transformation of N by soil microbial communities is directly tied to the pool of bioavailable N in soils (4, 5). Thus, understanding the controls of N metabolism in soil microbes is key to determining, and potentially managing (6), the cycling of N in soils. Although genes and regulatory mechanisms for microbial N cycling processes have long-been identified in laboratory studies (7–9), the short-term dynamics and controls of N cycling in complex soil communities remain poorly understood. The availability of shotgun sequencing technologies to analyze microbial functioning in soil communities provides an opportunity to enhance our understanding of microbially mediated soil N cycling.

Measuring short-term responses of soil microbial populations to changes in the environment is crucial in understanding the role of microbes in biogeochemical cycling. Most biogeochemical transformations occur during short periods of intense microbial activity, when the active fraction of microbes may be up to 20 times higher than in bulk soil (10). This stimulation is often the result of a localized increase in nutrient concentrations, such as in the rhizosphere or an area of fresh organic matter decomposition. Despite the importance of these “hot moments”, only a few studies (e.g. 11, 12) have tracked changes in N-cycling gene transcription in soils.

Notably, the short-term (hours to days) transcriptional response of N-cycling genes in response to labile C inputs has yet to be determined. Microbial communities experience sudden changes in C and N availability associated with plant root exudation (13), trophic interactions (14, 15), and litter leachate (16). Since soil microbes are typically limited by labile C and energy (17–19), the addition of a C-rich substrate is expected to stimulate growth and activity (20), increasing the

demand for N (21). Whether N is derived from the uptake of organic N present in the substrate or mineral N available in the soil depends largely on the C:N of the substrate (22). For example, in Yang et al. 2016 (23) soil microbial communities assimilated organic N during the mineralization of added glycine, but in the presence of glucose the mineralization of glycine was initially suppressed and ammonium served as the main source of N. Simple sugars such as glucose have accordingly been shown to influence protease activity (24). The metabolic pathways for N immobilization have been well characterized *in vitro* (25). A majority of N assimilation into biomass occurs through the conversion of NH_4^+ into the amino acids glutamine and glutamate, which are used as sources of N for all other amino acids. Under low-to-moderate intracellular concentrations of NH_4^+, the enzymes glutamine synthetase (GS; encoded by *glnA*) and glutamate synthase (GOGAT; *gltS*) convert NH_4^+ to glutamate in a two-step reaction referred to as the GS-GOGAT pathway (26). Under high concentrations of NH_4^+, the enzyme glutamate dehydrogenase (GDH; *gudB, gdhA*) converts NH_4^+ directly to glutamate in a one-step reversible reaction (27).

Since both the GS-GOGAT pathway and GDH require N as NH_4^+, other forms of inorganic N must be converted to ammonium before conversion into biomass. In the case of nitrate and nitrite, the reduction to ammonium occurs through either assimilatory nitrate reduction or, under anoxic conditions, dissimilatory nitrate reduction to ammonium (DNRA; Table S1) (28). The conversion of atmospheric N_2 to ammonium by diazotrophs is catalyzed by the enzyme nitrogenase (*nifDHK*) (29).

The mechanisms regulating N uptake in response to C have been extensively studied *in vitro* (8, 25). The complex regulatory network includes a specialized sigma factor (σ^{54}; *rpoN*), three transcriptional regulators, and a phosphorylation cascade comprised of post-modification

enzymes, PII proteins, and a two-component regulator (30). The activity of many of the enzymes and proteins in the phosphorylation cascade is tightly controlled by cellular concentrations of glutamine and oxoglutarate (31). Since the concentration of oxoglutarate is impacted by the activity of the TCA cycle, the regulation of N cycling is directly tied to C metabolism (32). Carbon substrate addition is also thought to influence dissimilatory N cycling processes such as nitrification and denitrification. In nitrification, ammonia is oxidized to nitrite and then nitrate. Often the steps of this process occur in different organisms (33), however complete ammonia oxidizers have also been described (34, 35).". In denitrification, nitrate is reduced to nitrite, nitric oxide, and then nitrous oxide and N_2. Nitrification and denitrification, beyond their ability to draw from the pools of ammonium and nitrate, also represent important avenues of inorganic N loss from soils via nitrate leaching and the release of N_2 and nitrous oxide, a potent greenhouse gas (36). The addition of glucose is expected to have both positive and negative effects on nitrification. Rates of autotrophic nitrification tend to decrease as heterotrophs outcompete autotrophic nitrifiers for ammonium (37), but rates of heterotrophic nitrification may increase after labile C inputs (38). Denitrification is more directly influenced by C availability and quality (39), and the abundance of mRNA transcripts associated with denitrification was stimulated with the addition of glucose in anoxic soil microcosms (40).

Despite our knowledge of the mechanisms and controls of N cycling and N metabolism, we do not yet fully understand how these genes are regulated within complex soil microbial communities. Metatranscriptomics allows us to capture the transcriptional profile of a microbial community, providing insight into the potential activity of a community at a given moment in time (41–43). Many studies utilizing this technique have focused on the influence of ecosystem level characteristics/properties on transcription, such as land-use, above ground cover,

seasonality, and climate (e.g. (44–49)). Although these studies contribute greatly to our understanding of community gene transcription, there is additional need to study the dynamic short-term responses of microbial communities to changes in C and N availability (50).

In order to fill this knowledge gap, we conducted a soil incubation study where we induced rapid immobilization of inorganic N by adding glucose. We selected glucose as it is a form of labile C commonly found in soils (51), and has been widely used to alleviate C limitation in soil microbial communities as a means to study growth (52, 53) and metabolic activity (50). We analyzed metagenomes and metatranscriptomes of the soil microbial community using high throughput shotgun sequencing to identify the response of N cycling genes over a 48-hour period. We hypothesized that the abundance of N-cycling genes in the metagenomes would not significantly change throughout the course of the 48-hour incubation, but that changes in activity would be immediately detected in the metatranscriptomes. We further hypothesized that there would be an upregulation of genes associated with inorganic N transport, N assimilation into biomass, and N metabolism regulation in response to labile C inputs, and that the abundance of these transcripts would track the concentrations of inorganic N. This work provides an in-depth look at the short-term transcriptional response of soil microbial communities during a central biogeochemical process in soils.

METHODS

Soil Sampling and Site Description

Soils were collected in the fall of 2017 from a long-term crop rotation experiment at the West Virginia University Certified Organic Farm near Morgantown, West Virginia, USA (39.647502° N, 79.93691° W; 243.8 – 475.2 m a.s.l.) (54, 55). Samples were taken from plots subject to a four-year conventionally tilled crop cycle consisting of corn, soybean, wheat and a mix of kale and cowpea. Manure was added every two years (during corn and wheat planting), and rye-vetch was added as a winter cover crop before replanting corn in the spring. From each plot, 10 cores 0-10 cm depth were collected and pooled.

Laboratory Incubation

Soil samples were shipped on ice to Northern Arizona University in Flagstaff, Arizona, USA. Soils from 3 plots were pooled, cleaned of roots and large debris, passed through a 2 mm sieve, and distributed between 64 glass Mason jars (500 mL), generating microcosms containing 30 g of soil each. The soil was preincubated at lab temperature (~ 23 °C) for 2 weeks prior to the glucose addition.

The microcosms received 1.6 mL of 0.13 M glucose solution, which added 0.7 mg of glucose C g^{-1} dry soil and raised the moisture content to 60% water holding capacity. Concentrations of glucose in this range have been demonstrated to stimulate soil microbial communities without creating a detrimental increase in osmotic pressure (52). Moreover, a brief trial incubation was conducted to ensure that this concentration of glucose would stimulate CO_2 production. Soils were incubated at lab temperature (~ 23 °C) under ambient lighting, but never direct sunlight. Every 4 hours, over a 48 h period, 5 jars were randomly selected and

destructively sampled. From each jar, we measured headspace CO_2 concentration, concentrations of NO_3^- and NH_4^+, and microbial biomass. A portion of each sample was immediately frozen using liquid N_2 and stored at -80°C for DNA and RNA extraction.

Since the addition of water may stimulate community activity and respiration, especially when starting with very dry soil (56, 57), we measured respiration in a parallel incubation wherein the same volume of water was added without glucose. Headspace CO_2 from these jars was measured and compared against the glucose additions in order to determine the overall effect of glucose and water on microbial respiration.

Biogeochemical Measurements and Analysis

To measure soil NO_3^- and NH_4^+ concentration, 8 g of soil from each destructively sampled jar were added to 40 ml of 1 M KCl solution, shaken for 1 hour, and filtered through Whatman no. 1 filter paper. Extracts were analyzed on a SmartChem 200 Discrete Analyzer (Westco Scientific Instruments, Brookfield, Connecticut, USA). Microbial biomass was measured using an extraction-fumigation-extraction technique (58), consisting of a 0.5 M K_2SO_4 extraction followed by a subsequent K_2SO_4 extraction with the addition of chloroform. The first extraction provided an estimate of the K_2SO_4 extractable organic C and N from each sample, while the second extraction provided an estimate of microbial biomass C (MBC) and N (MBN). Concentrations of extractable organic C and N were measured on a TOC-L (Shimadzu Corp, Kyoto, Japan). The concentration of CO_2 from the headspace of each microcosm was measured using a LI-6262 CO_2/H_2O Analyzer (Licor Industries, Omaha, Nebraska, USA) as described in Dijkstra et al. (2011) (59).

DNA and RNA Extraction and Sequencing

We extracted DNA and RNA just before (t_0) and 8 (t_8), 24 (t_{24}), and 48 (t_{48}) h after glucose addition (n=4). DNA and RNA were extracted using the RNeasy Powersoil Total RNA Kit (Qiagen) according to manufacturer instructions. DNA was separated from RNA using the RNeasy PowerSoil DNA Elution Kit (Qiagen). RNA samples were treated with RNase-Free DNase Set (Qiagen) to remove any DNA. Nucleic acid concentrations were determined with a Qubit fluorometer (Invitrogen, Carlsbad, California, USA), and purity was assessed with a NanoDrop ND-1000 spectrophotometer (Nanodrop Technologies, Wilmington, Delaware, USA). High-quality samples were sent to the Joint Genome Institute (JGI) for sequencing (60). Paired-end, 2 x 151 bp, libraries were prepared using the Illumina NovaSeq platform (Illumina Inc., San Diego, California, USA). Raw sequence reads were uploaded to the JGI genome portal (https://genome.jgi.doe.gov/portal/) under GOLD project ID Gs0135756. A more detailed description of the sequencing can be found in the data release (61).

Metagenome and Metatranscriptomic Analysis

Metatransciptomes were assembled by JGI using MEGAHIT v1.1.2 (62) (parameters "megahit ––k–list 23,43,63,83,103,123 —continue –o out.megahit") and metagenomes were assembled using SPAdes version 3.13.0 (63). Assembled metatranscriptomes and metagenomes were uploaded to the Integrated Microbial Genomes and Microbiomes (IMG/M) (64) pipeline for annotation. Full details of the bioinformatics pipeline, as well as SRA reference numbers can be found in the data release (61). From IMG/M we retrieved the number of reads for all genes attributed to functional orthologs in the Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology database (65), as well as taxonomic annotations against the IMG database. Contigs are

available through the JGI genome portal, and taxonomic and functional annotations of these contigs are available on the IMG/M database (http://img.jgi.doe.gov), under GOLD project ID Gs0135756. JGI Genome ID's for each sample, as well as sample metadata, can be found in Chuckran et al (2020; 61).

Normalization of KEGG functional annotations was preformed using the Bioconductor (66) program DESeq2 (67) in R. DESeq2 uses a negative binomial distribution to normalize read counts and estimates average $\log_2$ fold change (LFC) between harvests. Significant LFCs for each KEGG functional gene and transcript were determined through both a likelihood ratio test (for overall significance) and Wald test (for specific contrasts between timepoints) provided in DESeq2. Significance for both tests were assumed as a false discover rate (FDR) < 0.01. Prior to analysis, genes with less than 60 reads summed over all samples were discarded in an effort to reduce the FDR correction and improve detection of significant LFCs (68).

To assess differences in genes and transcripts composition over time, we performed permutational multivariate analysis of variance (PERMANOVA) on our metagenomes and metatranscriptomes. PERMANOVAs were conducted using Bray-Curtis dissimilarity matrices of the square root transformed normalized read counts with 999 permutations. A SIMPER analysis was used to determine genes which most strongly influenced differences between harvests. PERMANOVAs and SIMPER analyses were conducted using the vegan package (69) for R.

To assess the response of N metabolism to the addition of glucose, KEGG Orthology identifiers (K numbers) were grouped according to KEGG pathways and modules associated with N cycling (70), and K numbers representing regulatory genes controlling N metabolism were identified (8, 25) (Table S1). The response of C metabolism was determined by grouping K numbers by KEGG modules associated with glucose uptake, specifically the Entner-Doudoroff

pathway (KEGG module M0008), TCA cycle (M00009), pentose phosphate pathway (M00004), gluconeogenesis (M00003), and Glycolysis (M00001). From the TCA cycle, we also determined the response of isocitrate dehydrogenase (*icd*), which produces oxoglutarate, an important metabolite linking C and N metabolism (32). Counts and LFCs for K numbers were then averaged for each module to assess the overall response for each process. Results were visualized using the ggplot2 package (71) in R v 3.6.1 (72).

RESULTS

Biogeochemical Measurements

The concentration of NO_3^- decreased in the 24 hours after glucose addition and remained low for the remainder of the incubation (Fig. 1A). The concentration of NH_4^+ also decreased during the first 24 hours of the incubation and increased thereafter (Fig. 1B). Rates of CO_2 production increased from 4-16 hours and then decreased from 28-48 hours in response to glucose (Fig. 1C). We found that the addition of water only slightly influenced CO_2 production (Fig. S1), indicating that the majority of the stimulation was due to the addition of labile C. K_2SO_4 extractable organic carbon decreased for the first 20 hours and plateaued thereafter (Fig. 1D). Based on these biogeochemical measurements, we selected 4 timepoints (t_0, t_8, t_{24}, and t_{48}) from which we extracted DNA and RNA. These timepoints captured distinct phases of C and N availability that enabled us to test our hypotheses.

Microbial biomass C (MBC) moderately decreased throughout the incubation (Fig. S2A) and microbial biomass N (MBN) remained constant (Fig. S2B). Bacteria may exhibit some stoichiometric plasticity in response to nutrient inputs (73), however a decrease in biomass C:N in response to C inputs is counter-intuitive. Since the method of microbial biomass extraction used involves two extractions on the same sample (one before and after fumigation), incomplete extraction of the added glucose in the first extraction could result in an artificially high estimate of biomass C. We believe that it is far more likely that microbial biomass and stoichiometry did not change, and that changes in estimated MBC are likely the result of unextracted glucose remaining from the initial K_2SO_4 extraction.

Metagenomic and Metatranscriptomic Assembly and Annotation

Out of 16 soil samples from which DNA and RNA were extracted, 12 were successfully sequenced and assembled for metagenomic analysis and all 16 for metatranscriptomic analysis. For the metagenomes, the proportion of genes successfully annotated against the KEGG database varied from 23.4% to 25.6% of all genes per sample. Of the 6,876 functional KEGG orthologs identified in the metagenome analysis, 671 genes were in higher abundance while 332 were present in lower abundance (FDR < 0.01) after the addition of glucose. Glucose caused a shift in the relative abundance of functional genes (PERMANOVA, $F_{3.11} = 3.24$, $P < 0.01$; Fig. 2A). The genes that were most different in gene abundance relative to t_0 varied for each timepoint (SIMPER analysis; Table S3A), and not one of these genes was directly related to N uptake. Among these were the subunits of RNA polymerase *rpoB* and *rpoC*, which were in slightly lower abundance at t_8 (LFC -0.1, FDR > 0.1), and the regulatory gene for the lac operon, *lacI*, which was in a greater abundance at t_{24} and t_{48} (LFC 0.7, FDR < 0.01). The largest changes were found at t_{24} for low-abundant spore gemination proteins (Table S3B), specifically *gerKC* (KO6297) and *yfkQ* (K06307) which were 8.8 and 7.4 LFC more abundant than at t_0.
The proportion of transcripts successfully annotated against the KEGG database varied between 12.6% and 32% of all transcripts in a metatranscriptome. Transcripts for 5,448 functional genes were identified, of which 1,141 increased and 855 decreased in response to glucose. A PERMANOVA indicated significant shifts in the abundance of transcripts between timepoints ($F_{3, 15} = 8.07$, $P < 0.01$; Fig. 2B). Transcripts encoding for *amt* and *glnA* contributed the most to dissimilarity with t_0 (SIMPER analysis), combined they explained 1% of the differences at t_8, 1% of differences at t_{24}, and 0.9% of differences at t_{48}.

Gene and Transcript Abundance of Nitrogen Cycling Processes

The abundance of N cycling genes was generally stable over time (Fig. 3A), with changes in gene abundance often being several orders of magnitude smaller than changes in transcript abundances. For metatranscriptomes, many genes associated with N uptake were highly upregulated in response to glucose (Fig. 3). Expression of genes encoding the GS-GOGAT pathway (GS - *glnA;* GOGAT - *gltS, gltD, gltB*) was consistently upregulated after glucose addition (FDR < 0.01), peaking at 8 h (Fig. 3B, Table S2). We did not find a similar trend for transcripts associated with glutamate dehydrogenase (GDH: *gudB, gdhA*). Instead we found variable increases and decreases in the expression for these genes which corresponded with different classes of GDH enzymes (Fig. 3B Table S2). In prokaryotes, GDH often uses NADH (EC 1.4.1.2), NADPH (EC 1.4.1.4) as cofactors, while GDH in eukaryotes can use both (NAD(P)H; EC 1.4.1.3) (74). Transcription of genes for EC 1.4.1.4 significantly increased early (t_8, LFC 1.542 ± 0.312; FDR < 0.01), and transcription for EC 1.4.1.2 trended higher later (t_{48}, LFC 2.229 ± 0.884; FDR < 0.1). The eukaryotic EC 1.4.1.2 gene GDH2 (K15371) was upregulated at t_{24} (LFC 1.350 ± 0.434; Table S2; FDR < 0.01) and EC 1.4.1.3 was slightly downregulated throughout (significantly at t_8, FDR < 0.01).

The abundance of transcripts encoding the ammonium transporter AmtB (*amt*) was significantly (FDR < 0.01) higher after glucose addition throughout the 48-h incubation (Fig. 3B, Table S2), peaking at t_8, where it was 16-fold higher than at t_0 (41,366 transcripts at t_8 vs 2,539 at t_0). A similar upregulation was found for genes associated with nitrate and nitrite transport across the membrane – 1500-fold increases compared to t_0 (from 2.6 to almost 2800 transcripts per sample at t_{24}; Fig. 3B).

Genes associated with assimilatory nitrate reduction (Fig. 3; Table S2) were strongly upregulated at t_8 and remained upregulated over the 48 h incubation period. In contrast, we found variable

responses of genes associated with DNRA. Most genes associated with the dissimilatory reduction of nitrate to nitrite were downregulated or not significantly affected, with a few exceptions. Nitrate reductase gamma subunits (*narI/narV*) were upregulated at t_{24} and t_{48}, and the genes *nirB* and *nirD*, which encode the small and large subunit of the cytosolic enzyme nitrite reductase, were significantly (FDR < 0.01) upregulated throughout the incubation (LFC 6.18 to 7.70; Fig. 3B). In contrast to these enzymes, abundance of transcripts that encode a periplasmic cytochrome c nitrite reductase (*nrfA* and *nrfH*) did not significantly change in response to C amendment.

Expression of all genes involved with nitrification were downregulated in response to glucose, and a majority of those genes (5 of 6) were significantly (FDR < 0.01) downregulated at some point during the incubation (Fig. 3B). Similarly, expression for most denitrification genes were downregulated throughout the incubation, with the exception of *narI* and *narV*, which encode for gamma subunits of nitrate reductase.

Transcripts for three genes that encode subunits of nitrogenase (*nifK, nifD,* and *nifH*) were detected, all of which were at very low abundance at t_0, t_8, and t_{48}. Only at t_{24} did we observe a strong significant (FDR < 0.01) upregulation for all 3 genes, up to 410-fold higher than t_0 for *nifH* (798 transcripts at t_{24} vs 1 at t_0; Fig. 3B).

We found that the vast majority of N cycling gene transcription could be attributed to bacteria and archaea (Fig. 4). Dissimilatory processes were largely from *Thaumarchaeota* and *Nitrospirae*, while assimilatory processes tended to be represented by *Proteobacteria, Actinobateria, and Acidobacteria*. Nitrogen fixation was heavily dominated by *Proteobacteria* (Fig. 4).

Regulation of N Cycling Genes

Generally, transcripts of genes associated with regulation of N metabolism increased after glucose addition (Fig. S3; Fig. 5). The abundance of ATase and UTase (*glnD* and *glnE)*, used for post-modification of glutamine synthetase (GS) and regulatory PII proteins respectively, initially increased at t_8 (2.18 ± 0.41 LFC and 4.31 ± 0.36 LFC; FDR < 0.01; Fig. S3; Fig. 5). UTase (*glnD*) but not ATase (glnE), continued to be significantly upregulated at t_{24} (3.79 ± 0.36 LFC) and t_{48} (2.75 ± 0.36 LFC; Fig. S3). Similar upregulation was noted for PII proteins GlnB (*glnB*; LFC > 2.9; FDR < 0.01; Fig. S3) and GlnK (*glnK*; LFC > 3.9; FDR < 0.01; Fig. S3), and the NtrC family genes *glnL* (FDR < 0.01) and *glnG* (FDR < 0.01 at t_8 and t_{24}; Fig. S3). No significant changes in transcript abundances were found for the transcriptional regulators *nac and lrp,* while *crp* and *rpoN* were slightly downregulated (LFC < -1) at t_8 and t_{24} (FDR < 0.01; Fig. S3; Fig. 5).

C Metabolism

The LFC and total number of normalized transcripts for processes involved with glucose breakdown (KEGG modules M00001, M00003, M00004, M0008, and M00009). increased from t_0 to t_8 and t_{24} (Fig. S4; Table S4; Tukey's HSD $p < 0.05$). Significant changes in transcript abundance after glucose amendment were found for the Entner-Doudoroff pathway and TCA cycle, including the enzyme isocitrate dehydrogenase (*icd*) which produces oxoglutarate, a metabolite which directly connects C and N metabolism (Fig. 5; Fig. S4B, Table S4).

DISCUSSION

Over a period of 48-hours after glucose addition we observed a substantial decrease in K_2SO_4 extractable organic C, an increase in CO_2 production rate, and an increase in the abundance of transcripts for genes associated with glucose breakdown. These changes coincided with a decrease in inorganic N and an increase in the transcript abundance of genes involved with inorganic N uptake, assimilation, and N metabolism regulation. These results demonstrate that soil microbial communities respond to labile C not only by upregulating genes associated with C metabolism, but also by rapidly increasing the transcription of genes responsible for N acquisition. Further, we found that genes for several forms of N acquisition (e.g., N fixation, assimilatory nitrate reduction, ammonium transport) were differentially transcribed over the 48 h incubation, indicating that changes in multiple microbially mediated N transformations occur within this small temporal window.

Inorganic N Uptake and Assimilation

The GS-GOGAT pathway appeared to be the predominant pathway through which ammonium was assimilated into biomass. The other main avenue of ammonium assimilation into biomass, the enzyme GDH, did not show a similar increase in transcript abundance and the abundance of GDH transcripts was substantially smaller than that of GS-GOGAT. This suggests that GS-GOGAT may be the dominant pathway for assimilation of inorganic N in soil microbial communities responding to labile C inputs. This finding is consistent with the notion that GDH is most active when NH_4^+ concentration is high and availability of C is low (27). Assays from soil microbial communities have also shown that GS activity increases in response to higher C to N ratios whereas GDH activity decreases (75). Further, we found that regulation of GDH

transcription appeared to be gene specific, with transcription for EC 1.4.1.4 increasing early and EC 1.4.1.2 increasing late. These results nicely follow concentrations of NH_4^+, as NADPH specific enzymes (EC 1.4.1.4) are generally used for ammonium assimilation (76) whereas NADH specific enzymes (EC 1.4.1.2) are commonly used for breakdown of glutamate to ammonium (77). These findings highlight the potential utility of measuring GDH and GS-GOGAT gene transcription for tracking the C and N balance within microbial communities at a given moment in time, which could be a useful approach when, for example, assessing how specific land use practices influence microbial metabolism and N cycling.

Various mechanisms for transporting inorganic N across the cell membrane were upregulated in response to glucose inputs. Notably, the gene *amtB*, which encodes for the ammonium transporter AmtB, was the second most abundant upregulated gene during the incubation (behind *glnA*). Similarly, we observed an upregulation of genes associated with nitrate and nitrite transport (KEGG module M00615) and assimilatory nitrate reduction, which coincided with a precipitous drop in the concentration of NO_3^-. Most genes involved with DNRA were not differentially expressed, indicating that nitrate reduction was primarily occurring under aerobic conditions. A notable exception were the genes *nirB* and *nirD*, which encode for the cytosolic enzyme nitrite reductase NirBD (78), which has been shown to be active in aerobic soils (79, 80) and may function as the nitrite reductase in assimilatory nitrate reduction (81). Although the upregulation of N transport genes in response to glucose is certainly not novel (30), these results are the first demonstration of this response in a soil microbial community metatranscriptome. Further, these responses show the short timeframes (within 8 h) in which soil microbial communities can respond to changes in C and N availability.

The finding that glucose addition strongly upregulated genes encoding for nitrogenase, especially when NH_4^+ concentrations were low, is consistent with the idea that nitrogen fixation increases when N concentrations are low (82). N fixation has been shown to be activated by the addition of other limiting nutrients such as carbon or phosphorous (83, 84). We therefore believe that the upregulation of nitrogenase genes is a response to low concentrations of NH_4^+ and availability of labile C. The prompt upregulation, and subsequent downregulation, of nitrogenase genes also suggests that some portion of biological nitrogen fixation occurs rapidly in soils, or at the very least that the process is highly sensitive to concentrations of NH_4^+.

Connections Between C and N Metabolism

Interestingly, transcripts associated with NH_4^+ and NO_3^- transport maintained their high abundances despite concentrations of NO_3^- stabilizing and concentrations of NH_4 increasing (24-48 h into the incubation). One possible explanation is that the activity of these proteins is dictated through allosteric regulation which is tightly connected to the activity of both C and N metabolism (Fig. 5). For example, the ammonium transporter AmtB is allosterically inhibited by the PII protein GlnK which is indirectly controlled by internal concentrations of glutamine, an intermediate of N uptake through GS-GOGAT (Fig. 5), and oxoglutarate, an intermediate of the TCA cycle (Fig. 5; (32, 85)). In this way, internal concentrations of metabolites from both C and N metabolism may dictate N uptake.

The transcription of N regulatory genes reflects the importance of intermediate metabolites in regulation. We found that abundance of transcripts for transcriptional regulators (such as *nac, lrp,* and *crp*) and σ^{54} were either not affected or slightly reduced (Fig. 5). In contrast, transcripts for genes in the phosphorylation cascade, which links C and N metabolism through intermediate

metabolites, were more abundant after the addition of glucose (Fig. 5). The upregulation of the two component regulatory NtrB (*glnL*, *ntrB*) and NtrC (*glnG*, *ntrC*) within this cascade is especially noteworthy, as this system regulates ~75 genes associated with N acquisition, including glutamine synthetase (Fig. 5) (86).

Since the activity of this regulatory network is tightly controlled by internal concentrations of metabolites (30), it is not possible to determine the activity of many of these proteins through the metatranscriptome alone. However, it is noteworthy that almost all of the genes within this regulatory network were upregulated, even if the encoded protein potentially inhibited N transport or assimilation (e.g. GlnK; Fig. 5). This broad upregulation of genes in the phosphorylation cascade may be beneficial during C uptake, as it allows the concentration of nutrients and metabolites to control N uptake, thereby ensuring N uptake matches the supply of C (25, 32).

Nitrification and Denitrification

Most genes associated with nitrification and denitrification were significantly downregulated. Since nearly all nitrifiers in this soil were autotrophic archaea (55), this finding is consistent with the premise that addition of glucose reduces rates of autotrophic nitrification by reducing the amount of available ammonium (37). It is not especially surprising that we did not find an upregulation of denitrification genes, as denitrification is most prevalent in anoxic systems with high availabilities of nitrate.

Genetic Potential Versus Transcription

Notably, although we did observe a slight shift in the functional composition of our metagenomes, these changes did not track those found in the metatranscriptomes in either magnitude or direction. Changes contributing the most to dissimilarity tended to be slight shifts in highly abundant genes, such as *rpoB*, *rpoC*, and *lacI*. We found interesting differences in the abundance of spore forming proteins as nutrient availability declined, however since many of these proteins were uncommon and in low abundance, the chance of obtaining a false positive is much greater and we are therefore cautious to draw any conclusions based on these data alone. Changes in gene abundance for most N cycling genes were absent. These results suggest that understanding the response of soil microbial communities to short-term changes in the environment necessitates looking beyond the metagenome, as consequential microbial responses occur through changes in gene-expression. This is in line with other studies where the composition of transcripts shifts over hours or days (12, 87), whereas shifts in metagenomic community composition have been shown to occur after weeks or months (88) .

Our work represents a preliminary look into the short-term transcriptional response of microbial communities in response to a change in C availability, however there are a number of considerations moving forward. More work needs to be done focusing on this response in a variety soils, as nutrient availability and other soil properties will undoubtably influence this process. For example, soils high in C and low in N would likely not demonstrate a similar response as observed for this agricultural soil. Understanding how ecosystem properties influence the dynamics of transcriptional profiles is therefore necessary in determining short-term microbial contributions to biogeochemical cycling. Further, this work focused on a relatively short timeframe, however whether this increase in transcription persists or influences nutrient cycling on the scale of weeks to months remains to be seen. Finally, future efforts should

be made to observe these short-term effects *in situ*. Laboratory incubations are extremely useful for controlling environmental variables and isolating a particular response. However, it is likely that under field conditions, and in the presence of plant roots, factors other than C availability will affect the gene-expression at the same time and to different degrees, potential masking the response observed in this short-term laboratory experiment.

CONCLUSIONS

Our results indicate strong and rapid upregulation of genes associated with uptake of inorganic N, assimilatory nitrate and nitrite reduction, GS-GOGAT pathway, and the regulatory network underlying N cycling. Further, the majority of upregulation occurred in pathways which are largely aerobic and heterotrophic, suggesting that these processes dominate the short-term response to labile C in these soils. Perhaps most importantly, this work highlights the importance of microbial gene transcription in controlling short-term biogeochemical cycling in soils. Within the 48 h incubation we found that microbially mediated transformations of N were well reflected in the metatranscriptome but not in the metagenome or in microbial biomass. The short-term transcriptional responses of soil microbes may therefore serve an important role in determining how biogeochemical fluxes respond to immediate changes in the environment.

ACKNOWLEDGEMENTS

This work was supported by funding from the USDA National Institute of Food and Agriculture Foundational Program (award #2017-67019-26396) and additional support for PD was provided by the U.S. Department of Energy, Office of Biological and Environmental Research, Genomic Science Program LLNL 'Microbes Persist' Soil Microbiome Scientific Focus Area (award #SCW1632). The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH11231. We would like to thank Rebecca Mau, Michaela Hayer, Alicia Purcell, and Ayla Martinez for their assistance with laboratory analyses; Sam Bunkers, Kieston Guidry, and Kiara Nelson for their help downloading and cleaning the data; and Isaac Shaffer for his assistance with the analysis. We would also like to thank the Joint Genome Institute for their work in sequencing and assembly, specifically: Marcel Huntemann, Alicia Clum, Brian Foster, Bryce Foster, Simon Roux, Krishnaveni Palaniappan, Neha Varghese, Supratim Mukherjee, T.B.K. Reddy, Chris Daum, Alex Copeland, Natalia N. Ivanova, Nikos C. Kyrpides, Tijana Glavina del Rio, and Emiley A. Eloe-Fadrosh.

Competing interests:

The authors have no competing interests to disclose

REFERENCES

1. LeBauer DS, Treseder KK. 2008. Nitrogen limitation of net primary productivity in terrestrial ecosystems is globally distributed. Ecology 89:371–379.

2. Skiba U, Smith KA. 2000. The control of nitrous oxide emissions from agricultural and natural soils. Chemosph - Glob Chang Sci 2:379–386.

3. Camargo JA, Alonso Á. 2006. Ecological and toxicological effects of inorganic nitrogen pollution in aquatic ecosystems: A global assessment. Environ Int 32:831–849.

4. Mooshammer M, Wanek W, Hämmerle I, Fuchslueger L, Hofhansl F, Knoltsch A, Schnecker J, Takriti M, Watzka M, Wild B, Keiblinger KM, Zechmeister-Boltenstern S, Richter A. 2014. Adjustment of microbial nitrogen use efficiency to carbon:Nitrogen imbalances regulates soil nitrogen cycling. Nat Commun 5:1–7.

5. Hallin S, Jones CM, Schloter M, Philippot L. 2009. Relationship between n-cycling communities and ecosystem functioning in a 50-year-old fertilization experiment. ISME J 3:597–605.

6. Batista MB, Dixon R. 2019. Manipulating nitrogen regulation in diazotrophic bacteria for agronomic benefit. Biochem Soc Trans 47:603–614.

7. Marzluf GA. 1997. Genetic regulation of nitrogen metabolism in the fungi. Microbiol Mol Biol Rev 61:17–32.

8. Reitzer L. 2003. Nitrogen assimilation and global regulation in Escherichia coli. Annu Rev Microbiol 57:155–176.

9. Cebolla A, Palomares AJ. 1994. Genetic regulation of nitrogen fixation in Rhizobium meliloti. Microbiologia 10:371–384.

10. Kuzyakov Y, Blagodatskaya E. 2015. Microbial hotspots and hot moments in soil: Concept & review. Soil Biol Biochem 83:184–199.

11. Albright MBN, Johansen R, Lopez D, Gallegos-Graves LV, Steven B, Kuske CR, Dunbar J. 2018. Short-term transcriptional response of microbial communities to nitrogen fertilization in a pine forest soil. Appl Environ Microbiol 84:e00598-18.

12. León-Sobrino C, Ramond J-B, Maggs-Kölling G, Cowan DA. 2019. Nutrient acquisition, rather than stress response over diel cycles, drives microbial transcription in a hyper-arid Namib Desert soil. Front Microbiol 10:1054.

13. Coskun D, Britto DT, Shi W, Kronzucker HJ. 2017. How plant root exudates shape the nitrogen cycle. Trends Plant Sci 22:661–673.

14. Trap J, Bonkowski M, Plassard C, Villenave C, Blanchart E. 2016. Ecological importance of soil bacterivores for ecosystem functions. Plant Soil 398:1–24.

15. Trubl G, Jang H Bin, Roux S, Emerson JB, Solonenko N, Vik DR, Solden L, Ellenbogen J, Runyon AT, Bolduc B, Woodcroft BJ, Saleska SR, Tyson GW, Wrighton KC, Sullivan MB, Rich VI. 2018. Soil Viruses Are Underexplored Players in Ecosystem Carbon Processing. mSystems 3:1–21.

16. Kuzyakov Y. 2010. Priming effects: Interactions between living and dead organic matter. Soil Biol Biochem 42:1363–1371.

17. Demoling F, Figueroa D, Bååth E. 2007. Comparison of factors limiting bacterial growth in different soils. Soil Biol Biochem 39:2485–2495.

18. Hobbie JE, Hobbie EA. 2013. Microbes in nature are limited by carbon and energy: the starving-survival lifestyle in soil and consequences for estimating microbial rates. Front Microbiol 4:324.

19. Schimel JP, Weintraub MN. 2003. The implications of exoenzyme activity on microbial carbon and nitrogen limitation in soil: a theoretical model. Soil Biol Biochem 35:549–563.

20. Papp K, Hungate BA, Schwartz E. 2019. Glucose triggers strong taxon□specific responses in microbial growth and activity: insights from DNA and RNA qSIP. Ecology ecy.2887.

21. Kamble PN, Bååth E. 2014. Induced N-limitation of bacterial growth in soil: Effect of carbon loading and N status in soil. Soil Biol Biochem 74:11–20.

22. Geisseler D, Horwath WR, Joergensen RG, Ludwig B. 2010. Pathways of nitrogen utilization by soil microorganisms - A review. Soil Biol Biochem 42:2058–2067.

23. Yang L, Zhang L, Geisseler D, Wu Z, Gong P, Xue Y, Yu C, Juan Y, Horwath WR. 2016. Available C and N affect the utilization of glycine by soil microorganisms. Geoderma 283:32–38.

24. Geisseler D, Horwath WR. 2008. Regulation of extracellular protease activity in soil in response to different sources and concentrations of nitrogen and carbon. Soil Biol Biochem 40:3040–3048.

25. Chubukov V, Gerosa L, Kochanowski K, Sauer U. 2014. Coordination of microbial metabolism. Nat Rev Microbiol 12:327–340.

26. Yuan J, Doucette CD, Fowler WU, Feng X, Piazza M, Rabitz HA, Wingreen NS, Rabinowitz JD. 2009. Metabolomics□driven quantitative analysis of ammonia assimilation in *E. coli*. Mol Syst Biol 5:302.

27. Sharkey MA, Engel PC. 2008. Apparent negative co-operativity and substrate inhibition in overexpressed glutamate dehydrogenase from *Escherichia coli*. FEMS Microbiol Lett 281:132–139.

28. Lin JT, Stewart V. 1997. Nitrate assimilation by bacteria. Adv Microb Physiol 39:1–30.

29. Zehr JP, Turner PJ. 2001. Nitrogen fixation: Nitrogenase genes and gene expression. Methods Microbiol 30:271–286.

30. van Heeswijk WC, Westerhoff H V., Boogerd FC. 2013. Nitrogen assimilation in *Escherichia coli*: Putting molecular data into a systems perspective. Microbiol Mol Biol Rev 77:628–695.

31. Merrick MJ. 1993. In a class of its own — the RNA polymerase sigma factor σ;54 (σN). Mol Microbiol 10:903–909.

32. Huergo LF, Dixon R. 2015. The emergence of 2-oxoglutarate as a master regulator metabolite. Microbiol Mol Biol Rev 79:419–35.

33. Stein LY, Klotz MG. 2016. The nitrogen cycle. Curr Biol 26:R94–R98.

34. Daims H, Lebedeva E V., Pjevac P, Han P, Herbold C, Albertsen M, Jehmlich N, Palatinszky M, Vierheilig J, Bulaev A, Kirkegaard RH, Von Bergen M, Rattei T, Bendinger B, Nielsen PH, Wagner M. 2015. Complete nitrification by Nitrospira bacteria. Nature 528:504–509.

35. Van Kessel MAHJ, Speth DR, Albertsen M, Nielsen PH, Op Den Camp HJM, Kartal B, Jetten MSM, Lücker S. 2015. Complete nitrification by a single microorganism. Nature 528:555–559.

36. Hu H-W, Chen D, He J-Z. 2015. Microbial regulation of terrestrial nitrous oxide formation: understanding the biological pathways for prediction of emission rates. FEMS Microbiol Rev 021:729–749.

37. Verhagen FJM, Duyts H, Laanbroek HJ. 1992. Competition for ammonium between nitrifying and heterotrophic bacteria in continuously percolated soil columns. Appl Environ Microbiol 58:3303–3311.

38. Lan T, Liu R, Suter H, Deng O, Gao X, Luo L, Yuan S, Wang C, Chen D. 2020. Stimulation of heterotrophic nitrification and N2O production, inhibition of autotrophic nitrification in soil by adding readily degradable carbon. J Soils Sediments 20:81–90.

39. Tiedje JM, Sexstone AJ, Myrold DD, Robinson JA. 1983. Denitrification: ecological niches, competition and survival. Antonie Van Leeuwenhoek 48:569–583.

40. Henderson SL, Dandie CE, Patten CL, Zebarth BJ, Burton DL, Trevors JT, Goyer C. 2010. Changes in denitrifier abundance, denitrification gene mRNA levels, nitrous oxide emissions, and denitrification in anoxic soil microcosms amended with glucose and plant residues. Appl Environ Microbiol 76:2155–2164.

41. Carvalhais LC, Dennis PG, Tyson GW, Schenk PM. 2012. Application of metatranscriptomics to soil environments. J Microbiol Methods 91:246–251.

42. Moran MA. 2009. Metatranscriptomics: Eavesdropping on complex microbial communities. Microbe 4:329.

43. Helbling DE, Ackermann M, Fenner K, Kohler HPE, Johnson DR. 2012. The activity level of a microbial community function can be predicted from its metatranscriptome. ISME J 6:902–904.

44. Nacke H, Fischer C, Thürmer A, Meinicke P, Daniel R. 2014. Land use type significantly affects microbial gene transcription in soil. Microb Ecol 67:919–930.

45. Damon C, Lehembre F, Oger-Desfeux C, Luis P, Ranger J, Fraissinet-Tachet L, Marmeisse R. 2012. Metatranscriptomics reveals the diversity of genes expressed by eukaryotes in forest soils. PLoS One 7:e28967.

46. Žifčáková L, Větrovský T, Howe A, Baldrian P. 2016. Microbial activity in forest soil reflects the changes in ecosystem properties between summer and winter. Environ Microbiol 18:288–301.

47. Kim Y, Liesack W. 2015. Differential assemblage of functional units in paddy soil microbiomes. PLoS One 10:e0122221.

48. Bei Q, Moser G, Wu X, Müller C, Liesack W. 2019. Metatranscriptomics reveals climate change effects on the rhizosphere microbiomes in European grassland. Soil Biol Biochem 138:107604.

49. Baldrian P, Kolařík M, Štursová M, Kopecký J, Valášková V, Větrovský T, Žifčáková L, Šnajdr J, Rídl J, Vlček Č, Voříšková J. 2012. Active and total microbial communities in forest soil are largely different and highly stratified during decomposition. ISME J 6:248–258.

50. Anderson TH, Domsch KH. 1985. Maintenance carbon requirements of actively-metabolizing microbial populations under in situ conditions. Soil Biol Biochem 17:197–203.

51. Van Hees PAW, Jones DL, Finlay R, Godbold DL, Lundström US. 2005. The carbon we do not see - The impact of low molecular weight compounds on carbon dynamics and respiration in forest soils: A review. Soil Biol Biochem 37:1–13.

52. Reischke S, Rousk J, Bååth E. 2014. The effects of glucose loading rates on bacterial and fungal growth insoil. Soil Biol Biochem 70:88–95.

53. Reischke S, Kumar MGK, Bååth E. 2015. Threshold concentration of glucose for bacterial growth in soil. Soil Biol Biochem 80:218–223.

54. Pena-Yewtukhiw EM, Romano EL, Waterland NL, Grove JH. 2017. Soil health indicators during transition from row crops to grass–legume sod. Soil Sci Soc Am J 0:0.

55. Walkup J, Freedman Z, Kotcon J, Morrissey EM. 2020. Pasture in crop rotations influences microbial biodiversity and function reducing the potential for nitrogen loss from compost. Agric Ecosyst Environ 304.

56. Birch HF. 1958. The effect of soil drying on humus decomposition and nitrogen availability. Plant Soil 10:9–31.

57. Barnard RL, Blazewicz SJ, Firestone MK. 2020. Citation Classic Rewetting of soil: Revisiting the origin of soil CO 2 emissions. Soil Biol Biochem 147:107819.

58. Brookes PC, Landman A, Pruden G, Jenkinson DS. 1985. Chloroform fumigation and the release of soil nitrogen: A rapid direct extraction method to measure microbial biomass nitrogen in soil. Soil Biol Biochem 17:837–842.

59. Dijkstra P, Dalder JJ, Selmants PC, Hart SC, Koch GW, Schwartz E, Hungate BA. 2011. Modeling soil metabolic processes using isotopologue pairs of position-specific 13C-labeled glucose and pyruvate. Soil Biol Biochem 43:1848–1857.

60. Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, Smirnova T, Grigoriev I V., Dubchak I. 2014. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Res 42.

61. Chuckran PF, Huntemann M, Clum A, Foster B, Foster B, Roux S, Palaniappan K, Varghese N, Mukherjee S, Reddy TBK, Daum C, Copeland A, Ivanova NN, Kyrpides NC, del Rio TG, Eloe-Fadrosh EA, Morrissey EM, Schwartz E, Fofanov V, Hungate B, Dijkstra P. 2020. Metagenomes and Metatranscriptomes of a Glucose-Amended Agricultural Soil. Microbiol Resour Announc 9.

62. Li D, Liu CM, Luo R, Sadakane K, Lam TW. 2015. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31:1674–1676.

63. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A V., Sirotkin A V., Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477.

64. Chen I-MA, Chu K, Palaniappan K, Pillay M, Ratner A, Huang J, Huntemann M, Varghese N, White JR, Seshadri R, Smirnova T, Kirton E, Jungbluth SP, Woyke T, Eloe-

Fadrosh EA, Ivanova NN, Kyrpides NC. 2019. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. Nucleic Acids Res 47:D666–D677.

65. Kanehisa M, Goto S. 2000. Yeast Biochemical Pathways. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30.

66. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, MaCdonald J, Obenchain V, Oleš AK, Pagès H, Reyes A, Shannon P, Smyth GK, Tenenbaum D, Waldron L, Morgan M. 2015. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods 12:115–121.

67. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15.

68. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. 2016. A survey of best practices for RNA-seq data analysis. Genome Biol 17:13.

69. Oksanen AJ, Blanchet FG, Kindt R, Legen- P, Minchin PR, Hara RBO, Simpson GL, Solymos P, Stevens MHH. 2019. vegan: Community ecology package.

70. Kanehisa M, Sato Y. 2019. KEGG Mapper for inferring cellular functions from protein sequences. Protein Sci.

71. Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysisle. Springer-Verlag New York.

72. Team RC. 2018. R: A language and environment for statistical computing. R Found Stat Comput Vienna, Austria.

73. Mooshammer M, Wanek W, Zechmeister-Boltenstern S, Richter A. 2014. Stoichiometric imbalances between terrestrial decomposer communities and their resources: Mechanisms and implications of microbial adaptations to their resources. Front Microbiol 5:22.

74. Smith EL, Austen BM, Blumenthal KM, Nyc JF. 1975. Glutamate DehydrogenasesEnzymes 3rd ed. Academic Press.

75. Geisseler D, Doane TA, Horwath WR. 2009. Determining potential glutamine synthetase and glutamate dehydrogenase activity in soil. Soil Biol Biochem 41:1741–1749.

76. Duncan PA, White BA, Mackie RI. 1992. Purification and properties of NADP-dependent glutamate dehydrogenase from Ruminococcus flavefaciens FD-1. Appl Environ Microbiol 58:4032–4037.

77. Miller SM, Magasanik B. 1990. Role of NAD-linked glutamate dehydrogenase in nitrogen metabolism in Saccharomyces cerevisiae. J Bacteriol 172:4927–4935.

78. Cole J. 1996. Nitrate reduction to ammonia by enteric bacteria: redundancy, or a strategy for survival during oxygen starvation? FEMS Microbiol Lett 136:1–11.

79. Ruiz B, Le Scornet A, Sauviac L, Rémy A, Bruand C, Meilhoc E. 2019. The nitrate assimilatory pathway in Sinorhizobium meliloti: Contribution to NO production. Front Microbiol 10:1526.

80. Pathan SI, Větrovský T, Giagnoni L, Datta R, Baldrian P, Nannipieri P, Renella G. 2018. Microbial expression profiles in the rhizosphere of two maize lines differing in N use efficiency. Plant Soil 433:401–413.

81. Stolz JF, Basu P. 2002. Evolution of nitrate reductase: Molecular and structural variations on a common function. ChemBioChem 3:198–206.

82. Dixon R, Kahn D. 2004. Genetic regulation of biological nitrogen fixation. Nat Rev Microbiol 2:621–631.

83. Benner JW, Vitousek PM. 2007. Development of a diverse epiphyte community in response to phosphorus fertilization. Ecol Lett 10:628–636.

84. Vitousek PM, Menge DNL, Reed SC, Cleveland CC. 2013. Biological nitrogen fixation: Rates, patterns and ecological controls in terrestrial ecosystems. Philos Trans R Soc B Biol Sci 368:1–9.

85. Coutts G. 2002. Membrane sequestration of the signal transduction protein GlnK by the ammonium transporter AmtB. EMBO J 21:536–545.

86. Zimmer DP, Soupene E, Lee HL, Wendisch VF, Khodursky AB, Peter BJ, Bender RA, Kustu S. 2000. Nitrogen regulatory protein C-controlled genes of *Escherichia coli*: Scavenging as a defense against nitrogen limitation. Proc Natl Acad Sci U S A 97:14674–14679.

87. Nuccio EE, Starr E, Karaoz U, Brodie EL, Zhou J, Tringe SG, Malmstrom RR, Woyke T, Banfield JF, Firestone MK, Pett-Ridge J. 2020. Niche differentiation is spatially and temporally regulated in the rhizosphere. ISME J 14:999–1014.

88. Mau RL, Liu CM, Aziz M, Schwartz E, Dijkstra P, Marks JC, Price LB, Keim P, Hungate BA. 2015. Linking soil bacterial biodiversity and soil carbon stability. ISME J 9:1477–1480.

LIST OF FIGURES

Figure 1

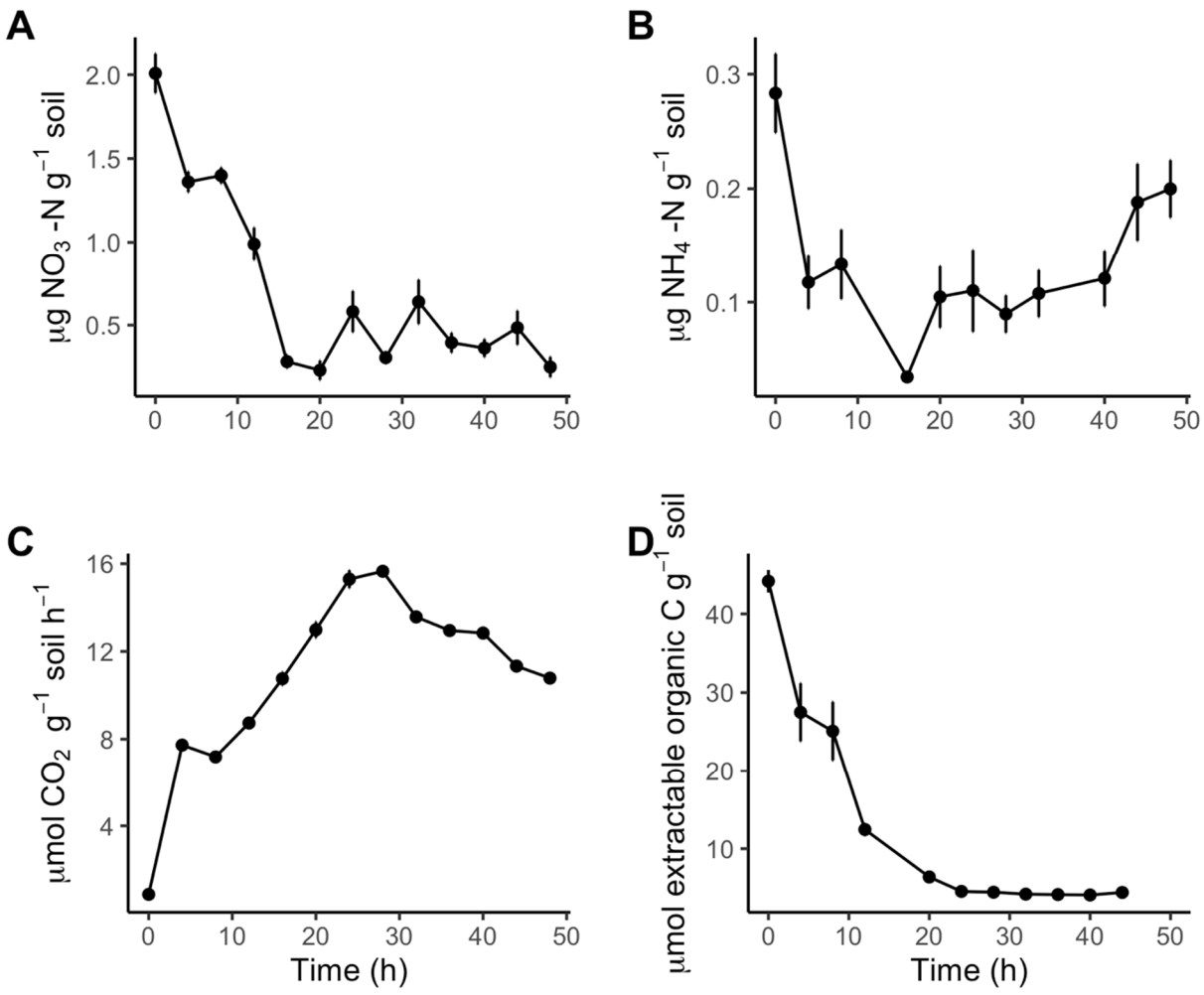

A
μg NO_3 -N g^{-1} soil
2.0
1.5
1.0
0.5
0
10
20
30
40
50
B
μg NH_4 -N g^{-1} soil
0.3
0.2
0.1
C
μmol CO_2 g^{-1} soil h^{-1}
16
12
8
4
Time (h)
D
μmol extractable organic C g^{-1} soil
40
30
20
10
Time (h)

Figure 2

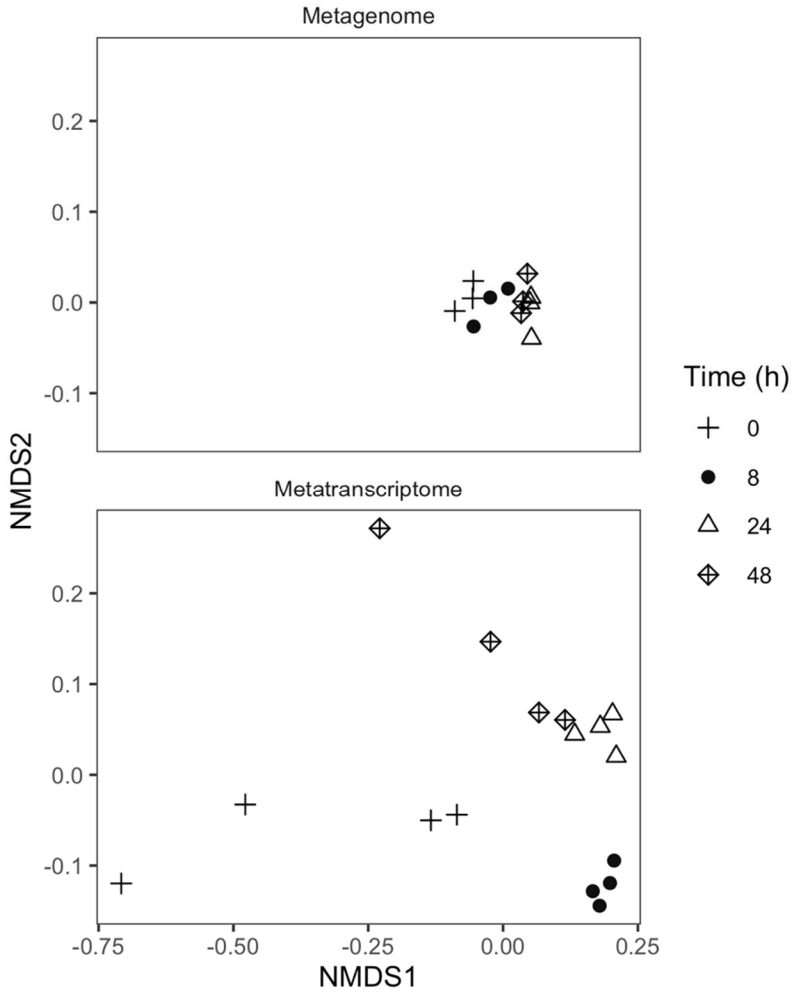

Metagenome
Metatranscriptome
NMDS2
NMDS1
0.2
0.1
0.0
-0.1
-0.75
-0.50
-0.25
0.00
0.25
Time (h)
0
8
24
48

Figure 3

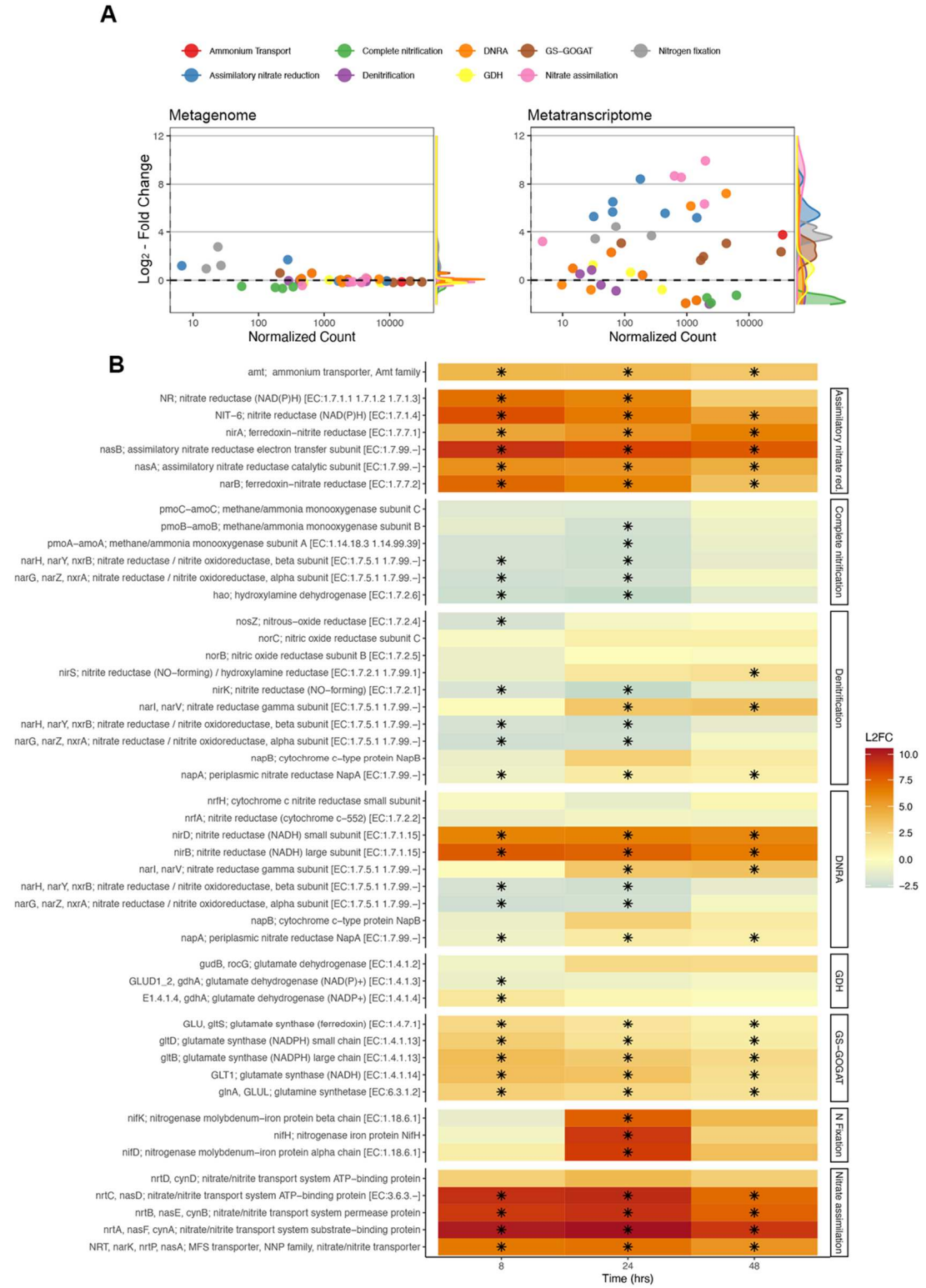

A
Ammonium Transport
Assimilatory nitrate reduction
Complete nitrification
Denitrification
DNRA
GDH
GS–GOGAT
Nitrate assimilation
Nitrogen fixation
Metagenome
Metatranscriptome
Log2 - Fold Change
Normalized Count
B
amt; ammonium transporter, Amt family
NR; nitrate reductase (NAD(P)H) [EC:1.7.1.1 1.7.1.2 1.7.1.3]
NIT–6; nitrite reductase (NAD(P)H) [EC:1.7.1.4]
nirA; ferredoxin–nitrite reductase [EC:1.7.7.1]
nasB; assimilatory nitrate reductase electron transfer subunit [EC:1.7.99.–]
nasA; assimilatory nitrate reductase catalytic subunit [EC:1.7.99.–]
narB; ferredoxin–nitrate reductase [EC:1.7.7.2]
pmoC–amoC; methane/ammonia monooxygenase subunit C
pmoB–amoB; methane/ammonia monooxygenase subunit B
pmoA–amoA; methane/ammonia monooxygenase subunit A [EC:1.14.18.3 1.14.99.39]
narH, narY, nxrB; nitrate reductase / nitrite oxidoreductase, beta subunit [EC:1.7.5.1 1.7.99.–]
narG, narZ, nxrA; nitrate reductase / nitrite oxidoreductase, alpha subunit [EC:1.7.5.1 1.7.99.–]
hao; hydroxylamine dehydrogenase [EC:1.7.2.6]
nosZ; nitrous–oxide reductase [EC:1.7.2.4]
norC; nitric oxide reductase subunit C
norB; nitric oxide reductase subunit B [EC:1.7.2.5]
nirS; nitrite reductase (NO–forming) / hydroxylamine reductase [EC:1.7.2.1 1.7.99.1]
nirK; nitrite reductase (NO–forming) [EC:1.7.2.1]
narI, narV; nitrate reductase gamma subunit [EC:1.7.5.1 1.7.99.–]
napB; cytochrome c–type protein NapB
napA; periplasmic nitrate reductase NapA [EC:1.7.99.–]
nrfH; cytochrome c nitrite reductase small subunit
nrfA; nitrite reductase (cytochrome c–552) [EC:1.7.2.2]
nirD; nitrite reductase (NADH) small subunit [EC:1.7.1.15]
nirB; nitrite reductase (NADH) large subunit [EC:1.7.1.15]
gudB, rocG; glutamate dehydrogenase [EC:1.4.1.2]
GLUD1_2, gdhA; glutamate dehydrogenase (NAD(P)+) [EC:1.4.1.3]
E1.4.1.4, gdhA; glutamate dehydrogenase (NADP+) [EC:1.4.1.4]
GLU, gltS; glutamate synthase (ferredoxin) [EC:1.4.7.1]
gltD; glutamate synthase (NADPH) small chain [EC:1.4.1.13]
gltB; glutamate synthase (NADPH) large chain [EC:1.4.1.13]
GLT1; glutamate synthase (NADH) [EC:1.4.1.14]
glnA, GLUL; glutamine synthetase [EC:6.3.1.2]
nifK; nitrogenase molybdenum–iron protein beta chain [EC:1.18.6.1]
nifH; nitrogenase iron protein NifH
nifD; nitrogenase molybdenum–iron protein alpha chain [EC:1.18.6.1]
nrtD, cynD; nitrate/nitrite transport system ATP–binding protein
nrtC, nasD; nitrate/nitrite transport system ATP–binding protein [EC:3.6.3.–]
nrtB, nasE, cynB; nitrate/nitrite transport system permease protein
nrtA, nasF, cynA; nitrate/nitrite transport system substrate–binding protein
NRT, narK, nrtP, nasA; MFS transporter, NNP family, nitrate/nitrite transporter
Assimilatory nitrate red.
Complete nitrification
Denitrification
DNRA
GDH
GS–GOGAT
N Fixation
Nitrate assimilation
L2FC
Time (hrs)

Figure 4

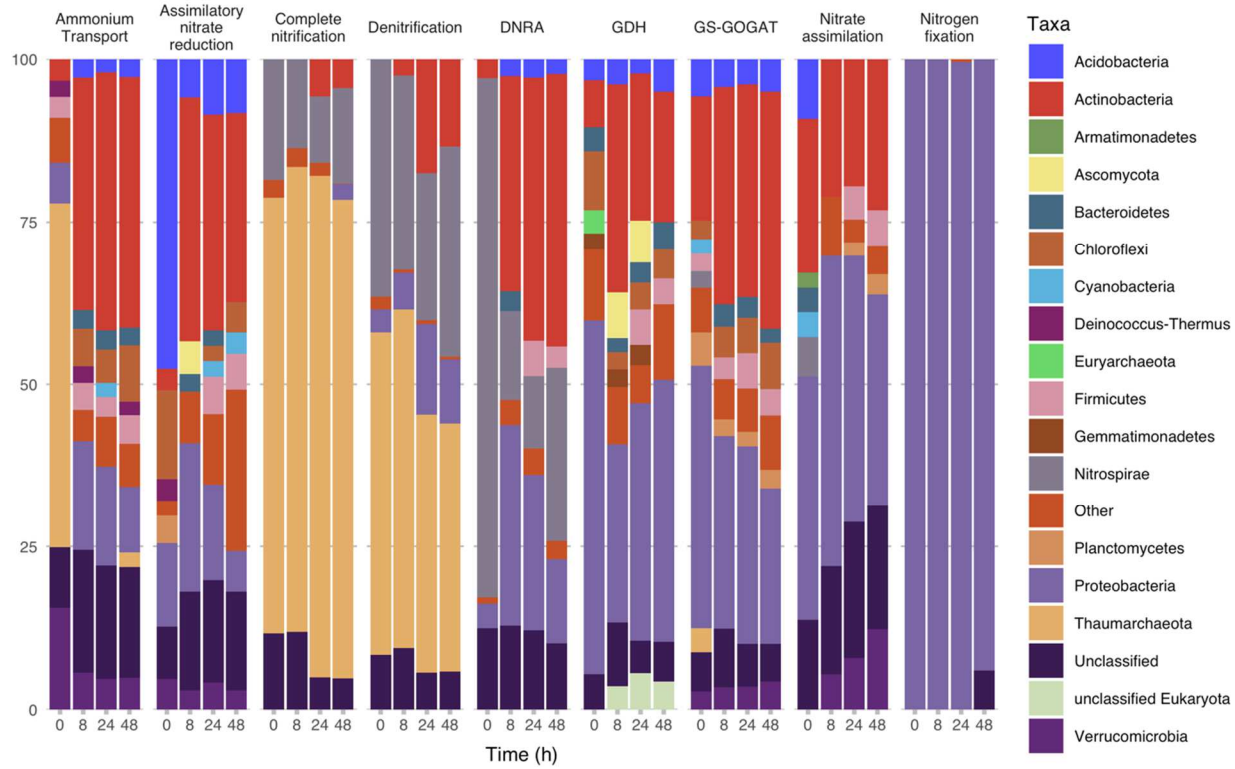
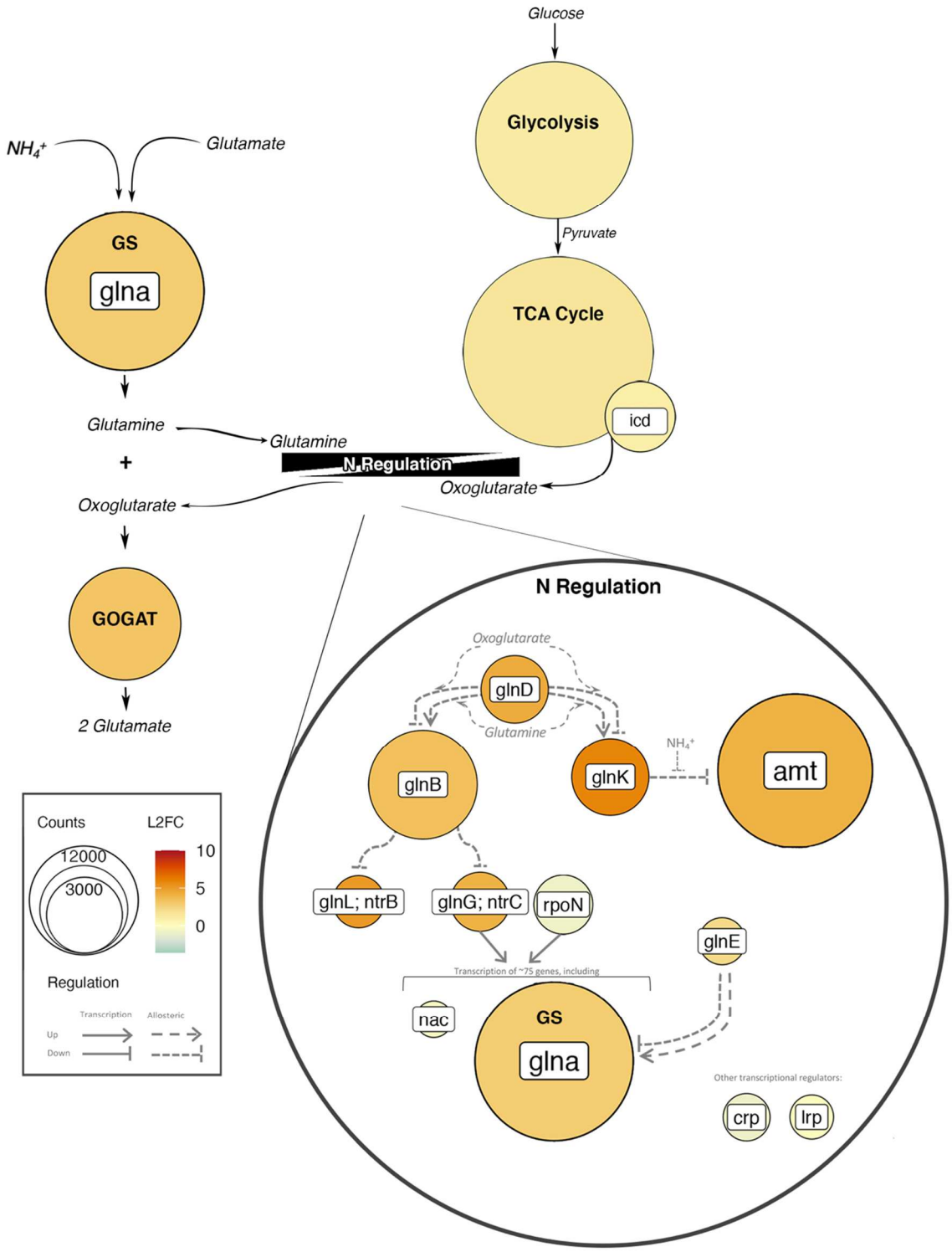

Ammonium Transport
Assimilatory nitrate reduction
Complete nitrification
Denitrification
DNRA
GDH
GS-GOGAT
Nitrate assimilation
Nitrogen fixation
100
75
50
25
0
0 8 24 48
Time (h)
Taxa
Acidobacteria
Actinobacteria
Armatimonadetes
Ascomycota
Bacteroidetes
Chloroflexi
Cyanobacteria
Deinococcus-Thermus
Euryarchaeota
Firmicutes
Gemmatimonadetes
Nitrospirae
Other
Planctomycetes
Proteobacteria
Thaumarchaeota
Unclassified
unclassified Eukaryota
Verrucomicrobia

Figure 5


Glucose
Glycolysis
Pyruvate
TCA Cycle
icd
Oxoglutarate
NH4+
Glutamate
GS
glna
Glutamine
Glutamine
+
N Regulation
Oxoglutarate
GOGAT
2 Glutamate
N Regulation
Oxoglutarate
glnD
Glutamine
glnB
glnK
NH4+
amt
glnL; ntrB
glnG; ntrC
rpoN
glnE
Transcription of ~75 genes, including
nac
GS
glna
Other transcriptional regulators:
crp
lrp
Counts
L2FC
12000
3000
10
5
0
Regulation
Transcription
Allosteric
Up
Down

CHAPTER 3

VARIATION IN GENOMIC TRAITS OF MICROBIAL COMMUNITIES AMONG ECOSYSTEMS

AUTHORS:

Peter F. Chuckran[1*]

Bruce A. Hungate[1]

Egbert Schwartz[1]

Paul Dijkstra[1]

[1] Center for Ecosystem Science and Society (ECOSS) and Department of Biological Sciences, Northern Arizona University, PO Box 5620, Flagstaff, AZ, USA

*Corresponding author: pfc25@nau.edu, pfchuckran@gmail.com

KEYWORDS:

Genome Size, GC content, Streamlining, Soil, Metagenomics, Sigma-factors

ABSTRACT

Free-living bacteria in nutrient limited environments often exhibit traits which may reduce the cost of reproduction, such as smaller genome size, low GC content, and fewer sigma (σ) factor and 16S rRNA gene copies. Despite the potential utility of these traits to detect relationships between microbial communities and ecosystem-scale properties, few studies have assessed these traits on a community-scale. Here, we analyzed these traits from publicly available metagenomes derived from marine, soil, host-associated, and thermophilic communities. In marine and thermophilic communities, genome size and GC content declined in parallel, consistent with genomic streamlining, with GC content in thermophilic communities generally higher than in marine systems. In contrast, soil communities averaging smaller genomes featured higher GC content and were often from low-carbon environments, suggesting unique selection pressures in soil bacteria. The abundance of specific σ-factors varied with average genome size and ecosystem type. In oceans, abundance of *fliA*, a σ-factor controlling flagella biosynthesis, was positively correlated with community average genome size – reflecting known trade-offs between nutrient conservation and chemotaxis. In soils, a high abundance of the stress response σ-factor gene *rpoS* was associated with smaller average genome size and often located in harsh and/or carbon-limited environments – a result which tracks features observed in culture and indicates an increased capacity for stress response in nutrient-poor soils. This work shows how ecosystem-specific constraints are associated with trade-offs which are embedded in the genomic features of bacteria in microbial communities, and which can be detected at the community level, highlighting the importance of genomic features in microbial community analysis.

INTRODUCTION

Assessing microbial communities through a trait-based framework highlights important relationships between microbes and their environment which may not be detectable through taxonomic analyses alone (Green, Bohannan and Whitaker 2008; Raes *et al.* 2011; Barberán *et al.* 2014; Fierer, Barberán and Laughlin 2014; Krause *et al.* 2014; Martiny *et al.* 2015). Notably, genomic characteristics such as genome size, GC content, number of regulatory genes, and number of 16S rRNA gene copies, have been shown to be indicators for growth rates (Vieira-Silva and Rocha 2010), life history strategies (Cobo-Simón and Tamames 2017) and population dynamics (Batut *et al.* 2014) of bacteria. Relationships between genomic features and environmental factors such as nutrient usage (Batut *et al.* 2014; Giovannoni, Cameron Thrash and Temperton 2014; Roller, Stoddard and Schmidt 2016), aboveground cover (Schmidt *et al.* 2018; Li *et al.* 2019), temperature (Sabath *et al.* 2013), and precipitation (Gravuer and Eskelinen 2017) have additionally demonstrated the potential utility of genomic traits for assessing the relationship between bacteria and their environment.

The genome size of free-living bacteria may be reduced by a process called genomic streamlining, wherein nutrient limitation selects for smaller genomes as a way to reduce the cost of reproduction (Giovannoni *et al.* 2005). Streamlined genomes are associated with a number of traits which also reduce reproductive costs, most notably a lower GC content (which reduces nitrogen requirements and is less costly to synthesize), fewer regulatory genes (specifically those encoding σ-factors), smaller intergenic spacer regions, and fewer 16S rRNA gene copies (Giovannoni, Cameron Thrash and Temperton 2014). Consequently, bacteria with streamlined genomes are thought to have a higher resource use efficiency and lower maximum growth rates compared to bacteria with larger genomes and more rRNA gene copies (Lauro *et al.* 2009),

although evidence for this relationship remains mixed (Klappenbach, Dunbar and Schmidt 2000; Vieira-Silva and Rocha 2010; Yooseph *et al.* 2010; Karcagi *et al.* 2016; Kirchman 2016; Kurokawa *et al.* 2016). Streamlining has long-been known to be highly prevalent in marine systems (Morris *et al.* 2002) where the streamlined SAR11 clade, with a genome of only ~1.3 Mbp, makes up 25% of all planktonic bacteria (Giovannoni 2017). As a result, much of the current knowledge regarding streamlining is based on marine systems, although the recently described streamlined (2.81 Mbp) Verrucomicrobia, *Candidatus Udaeobacter copiosus*, has been shown to be ubiquitous in soils, comprising up to 30% of recovered taxa in some grassland soils (Brewer *et al.* 2017)—indicating that genome reduction may also be an important force shaping soil bacteria.

Temperature can also influence genome size due to increased fitness of small cells at high temperatures (Sabath *et al.* 2013). Accordingly, small cells and smaller genomes are typically associated with higher optimal growth temperatures. This relationship is most pronounced in thermophilic communities (Wang, Cen and Zhao 2015), but has also been demonstrated in marine systems (Swan *et al.* 2013; Morán *et al.* 2015; Huete-Stauffer *et al.* 2016) and more recently in soils (Sorensen *et al.* 2019). These patterns between genome size, GC content, and number of 16S rRNA gene copies as a result of temperature-induced genome reduction often resemble patterns in streamlined genomes (Sabath *et al.* 2013).

Small genomes are also prevalent in host-associated bacteria. However, the processes underpinning the reduction in genome size involve several mechanisms, including drift, rapid mutation rate, or other mechanisms, which could be more important than streamlining (Batut *et al.* 2014). In environments where nutrients are abundant but population sizes small, deletions in bacterial genomes are more likely to become fixed in a population (Mira, Ochman and Moran

2001; Batut *et al.* 2014), a process particularly common in host-associated gut microbiota, where population sizes are small due to isolation (McCutcheon and Moran 2012). Bacteria subject to higher levels of mutation are more likely to be AT-rich since there is a mutational bias from GC → AT (Kuo, Moran and Ochman 2009; Hershberg and Petrov 2010; Hildebrand, Meyer and Eyre-Walker 2010; Batut *et al.* 2014). Since the mechanisms driving the evolution of host-associated bacteria often stray from streamlining, genome reduction in host-associated bacteria may yield different patterns in genome reduction. Specifically, streamlining, which is more a directional rather than stochastic process, will often select for specific genes (Batut *et al.* 2014). However, much of this knowledge concerning bacterial genomic traits has been derived from cultures or isolates. This presents substantial bias in our understanding of these relationships (Gweon, Bailey and Read 2017), especially for genomic traits of bacteria in complex microbial communities (Rinke *et al.* 2013), as most bacterial taxa have never been cultured or isolated. An alternative approach is to examine genomic traits on a community level *in situ*. By observing community-derived metrics of genomic traits we broaden our understanding of the distribution and implication of these traits as they occur in the natural world. This is an important practice for microbial ecology as there has been growing interest in trait dimensions which might improve our assessment of community function (analogous to those existing for plants; Westoby *et al.* 2021), yet little work has been done to observe these traits on the community level. Such metrics could be valuable in the comparison of communities across landscapes and ecosystems. Genomic traits such as GC content, number of regulatory genes, and average genome size may be especially useful for this purpose, as they can often be easily estimated from metagenomic datasets and do not require an extensive knowledge of the taxa within the community. The relative ease with which these traits may be derived makes them ideal metrics for large-scale

comparisons. This represents a potentially valuable tool for linking microbial communities with ecosystem-level processes.

The ability to leverage these traits to gain insight into function, assembly, or evolutionary relationships remains untested. A necessary step towards building a more comprehensive understanding of community-derived traits includes assessment of the distribution of these traits across systems, such has been done numerous times for isolates. Here we present a comparison of genomic traits from 116 metagenomes from soil, marine, host-associated, and thermophilic systems. These systems were chosen as they represent distinct environments which exert unique evolutionary pressures on genomic traits which might produce predictable outcomes: streamlining in oceans; temperature-induced genome reduction in thermophiles; drift in host-associated communities. Several mechanisms have been shown to influence genome size in soils; however, the predominant force is not well understood. Isolate genomes in soils tend to be comparatively larger than other systems (Sabath *et al.* 2013) which is thought to be a result of the increased metabolic diversity (Barberán *et al.* 2014). The overall aim of this study is to assess whether genomic traits measured at the community level track relationships which have been observed in isolates. Accordingly, we hypothesize that, consistent with trends in isolates, the average genome size in soil microbial communities will be larger than in marine, host-associated, or thermophilic communities. We also predict that GC content will be positively correlated with average genome size in free-living soil, marine, and thermophilic communities—consistent with trends from streamlined and thermophilic isolates. Finally, we predict that while both free-living and host-associated communities with small average genome sizes will demonstrate a low GC content, free-living communities will also exhibit additional streamlined traits such as a reduced number of σ-factor and rRNA gene copies.

MATERIALS AND METHODS

Dataset Curation

Metagenomes from soil, marine, thermophilic, and host-associated communities were downloaded from the Integrated Microbial Genomes & Microbiomes (IMG/M) (Chen *et al.* 2019) system. Data were used in accordance to JGI IMG/M data release policies (https://jgi.doe.gov/user-programs/pmo-overview/policies/) and studies were only used under the follow conditions: 1) The studies were previously published with a corresponding publication on the IMG database or; 2) We were granted written consent from the team which generated the data. This publication does not act as a primary publication for these studies and use of the data from the second group requires consent from the corresponding principal investigators of that study. We searched for soil and marine samples that were untreated and collected *in situ* systems (i.e. not an incubation or microcosm). If studies included any form of experimental manipulation, then only metagenomes from the control were selected. For thermophilic samples we searched for communities derived from natural hot-springs, and for host-associated samples we focused on animal-associated communities. We then selected samples which were both sequenced and assembled (MEGAHIT (Li *et al.* 2015) or SPAdes (Bankevich *et al.* 2012)) by the Joint Genome Institute (JGI) and where > 35 Mbp were assembled. Replicates appearing to be derived from a single sample (i.e. identical metadata and sample name) were discarded. In order to limit potential bias introduced by a specific study site or set of protocols of a given study, no more than 4 samples were used from any single geographical location and no more than 14 samples were selected from a single study. Ecosystem type was determined for soil samples using the available metadata and study description. In total, 116 samples from 30 different studies were used in this analysis (Supplemental Fig. 1; Supplemental Table 1&2).

Average genome size for each metagenome was estimated using the program MicrobeCensus (parameters -n 50000000) (Nayfach and Pollard 2015) on QC filtered reads accessed through the JGI Genome Portal (Nordberg *et al.* 2014). MicrobeCensus uses the abundance of single-copy genes to estimate the number of individuals in a population, which is then divided by the total number of read base-pairs to provide an estimate of the average genome size in a metagenome.

From IMG/M, we accessed the size of the metagenomic sample (bp), GC-%, total number of 16S rRNA gene copies, and the total number of σ factors identified by the KEGG Orthology database (Table 1; KEGG - Kanehisa and Goto 2000). We estimated the number of genomes per metagenome by dividing the total base pair count of the metagenome by the estimated average genome size from MicrobeCensus. The average number of 16S rRNA gene copies per genome and the number of σ-factors gene copies per genome was then determined by dividing the total number of 16S rRNA or σ-factor gene copies by the estimated number of genomes.

To ensure that any observed trends were not heavily influenced by the abundance of nonbacterial genomes, such as large eukaryotic genomes, we assessed the relationship between average genome size and the relative abundance of assembled bacterial reads. For each metagenome, we accessed the taxonomic assignments of mapped reads from IMG/M and then summed the total number of reads grouped by domain. The relationship between the relative abundance of bacteria and average genome size of the community was then calculated for each ecosystem to assign a cutoff which demonstrated the least amount of bias (as determined by linear regression). As a result, samples where bacteria made up less than < 95% of the assembled reads were discarded.

Since archaeal abundance in thermophilic microbial communities is often high, filtering samples with < 95% bacterial reads discarded a large number of thermophilic samples. Post filtering, only 5 thermophilic samples were left for analysis – a sample size ultimately too small to generate conclusions. Rather than omitting the thermophilic environments from our analysis entirely, and because small archaeal genomes abundance have been shown to be correlated with higher optimum growth temperatures (Sabath *et al.* 2013), we decided to include thermophilic samples with > 5 % archaeal abundance in several of the comparisons. Although these data do not examine bacterial streamlining specifically, we find that they still provide valuable insight into how genomic traits are distributed in these communities. Mixed thermophilic samples (those including > 5 % archaea) are shown separately in figures and analyses. In comparisons of genome size versus bacteria-specific traits, such as 16S rRNA gene copies or abundance of sigma factors, we only report samples where bacteria comprise > 95% of annotated reads.

Analysis

Multiple regression was used to determine the relationship between genome size and genomic characteristics – specifically, GC content, 16S rRNA gene relative abundance, the relative abundance of the total number of σ-factor genes, and the relative abundance of specific σ-factor genes as listed in Table 1. Models were constructed with the command lm or lmer from the R (v3.6.1 (Team 2018)) package lme4 (Bates *et al.* 2020). For each response variable, we constructed multiple models considering all parameters and interactions. Final models were selected using Akaike information criterion (AIC) values. The addition of a new parameter

resulting in a reduction of the AIC value by at least 4 indicated a significantly better fit with increased model complexity.

To assess the abundance of σ-factor genes between different ecosystems, we used both the multi-response permutation procedure (MRPP) as well as the permutational multivariate analysis of variance (PERMANOVA). The MRPP was conducted using all samples while PERMANOVA was conducted using 11 randomly selected genomes from each ecosystem to ensure balanced design. Both analyses were conducted using Bray-Curtis dissimilarity matrices constructed from the relative abundance of each σ-factor. To visualize differences in the distribution of different types σ-factors between ecosystems we used nonmetric multidimensional scaling (NMDS) on Bray-Curtis distances. MRPP, PERMANOVA and NMDS were done using the *vegan* package (Oksanen *et al.* 2019) in R (v3.6.1).

Isolates

To compare relationships between genomic characteristics of a microbial community with characteristics of isolates, we accessed over 6,000 isolates of bacteria, archaea, and fungi from the IMG/M system in June of 2020. Isolates were selected if they were (1) publicly available; (2) previously published; (3) sequenced by JGI. The associated publications for these isolates may be found in the Supplemental References. Metadata was used to group samples into one of three ecosystem types: soil, marine, thermophilic, or host-associated. To avoid potential bias introduced by large studies selecting for specific taxa, we randomly selected no more than 20 isolates from a single study. Relationships between genomic characteristics were analyzed using multiple regression analyses as described above for the analysis of community-level traits.

ANOVA was used to assess differences in the distribution of genomic characteristics between isolates and metagenomic averages.

RESULTS

Average Genome Size and GC Content

Average genome size was significantly different between ecosystems (ANOVA; $F_{4,111} = 135.9$, $p < 0.01$). Specifically, average genome size was higher in soils compared to marine, host-associated, or thermophilic communities (Fig. 1a, Tukey's HSD $p < 0.01$). GC content (%) varied between each ecosystem (ANOVA; $F_{4,111} = 140.3$, $p < 0.01$), and was highest in soil, followed by thermophilic, host-associated, and then marine communities (Fig. 1b). The relationship between GC content and average genome size varied between ecosystems (Fig. 1c). A comparison of multiple models, using AIC values as selection criteria, indicated that GC content was best predicted by average genome size, ecosystem, and their interaction ($F_{9,106} = 136.1$, $p < 0.01$, Supplemental Table 3). Specifically, GC content was positively correlated with average genome size in marine and thermophilic communities, negatively correlated in soil communities, and not significantly related in host-associated communities (Fig. 1c). The relationship between average genome size and GC content was offset between marine and thermophilic communities, wherein thermophilic communities had a higher GC content than marine communities with the same average genome size (Fig. 1c). The relationship between GC content and average genome size was strongly driven by the abundance of archaea in the mixed thermophilic samples (Supplemental Fig. 2). In soils, average genome size and GC content were significantly different between ecosystem types (ex. Deserts, grasslands, forests; ANOVA: Mbp - $F_{7,38} = 24.35$, $p < 0.01$; GC-% - $F_{7,38} = 4.986$, $p < 0.01$; Fig. 2).

The average genome size and GC content of the metagenomes fell within the range of isolates from each ecosystem (Fig. 3). However, the mean genome size and GC content derived

from metagenomes varied from isolates in both soil and thermophilic environments (ANOVA; $p < 0.05$), but not in marine environments.

16S rRNA gene copies and Sigma factors

Host-associated communities had the highest number of 16S rRNA gene copies per genome, followed by soils and then thermophilic and marine communities (Supplemental Fig. 3). A comparison of AIC values indicated that ecosystem type alone was the best predictor of 16S rRNA gene copies per genome (Supplemental Fig. 3, Supplemental Table 3).

The relative abundance of σ-factors genes per metagenome changed with estimates of average genome size and this relationship varied significantly between ecosystems (Fig. 4; Fig. 5a; Supplemental Table 3). Average genome size was significantly correlated with the relative abundance of σ-factors in thermophilic environments ($R^2 = 0.49$), but not in soil, marine, or host associated environments ($R^2 < 0.2$; Fig. 5a). The distribution of σ-factor types within a metagenome varied more between ecosystems than within (Fig. 4; Fig. 5b; MMRP, $A = 0.34$, $p < 0.01$), and ecosystems differed significantly (Fig. 4; Fig. 5b; PERMANOVA, $R^2 = 0.50$, $p < 0.01$).

The relationship between average genome size and the relative abundance of individual σ-factors was dependent on both ecosystem type and the type of σ-factor (Fig. 5c, Supplemental Table 4). In host-associated communities, the relative abundance of only one σ-factor, *sigH,* was significantly ($p= 0.018$) negatively correlated with average genome size. Abundance of all other sigma factors were unchanged with genome size in host-associated communities (Supplemental Table 4). In soil communities the relative abundance of *rpoH* per metagenome significantly increased ($p < 0.01$) with larger average genome size, while the relative abundance per

metagenome of *rpoS, sigH, sigB,* and *fliA* decreased ($p < 0.01$). In marine communities, we found that the relative abundance of *fliA*, *rpoE*, and *sigH* significantly increased ($p < 0.01$) with genome size, and the abundance of *rpoH*, and *rpoD* significantly decreased ($p < 0.01$). Due to the small samples size of thermophilic communities, we did not include the relationships between σ-factors and average genome size for thermophilic environments; however, correlation coefficients and statistics for all linear regressions between average genome size and σ-factor abundance for each ecosystem can be found in Supplemental Table 4. A visualization of average σ-factor copies per genome can be found in Supplemental Fig. 4.

DISCUSSION

The range of values for both genome size and GC content on the community level was substantially more narrow than those recorded for isolates, both from the literature (Sabath *et al.* 2013) and the IMG database. However, we did observe considerable variation both between and within different ecosystems. The observed within-ecosystem variation is likely a product of the range of ecosystems included in the analysis. For example, soil metagenomes were derived from deserts, grasslands, forests, tropical forests, and polar deserts, and traits accordingly tended to separate out by these habitats (Fig. 2). This work demonstrates the variability that exists within a specific ecosystem type and highlights the potential utility of genomic traits in studies comparing multiple habitat types. Between ecosystems, microbial communities in marine, host, and thermophilic environments had a smaller average genome size and lower GC content than those in soil, consistent with our first hypothesis based on previous findings from studies using bacterial isolates and single-amplified genomes (Raes *et al.* 2007; Giovannoni, Cameron Thrash and Temperton 2014; Cobo-Simón and Tamames 2017). Although small genomes may persist in soil communities, larger genomes tend to be more abundant (Barberán *et al.* 2014; Brewer *et al.* 2017); a feature often attributed to the advantage gained from the increased abundance of secondary metabolite genes in large soil genomes (Konstantinidis and Tiedje 2004).

Since smaller genomes tend to have lower GC content (Bentley and Parkhill 2004), we expected to find a positive correlation between GC content and average genome size for each ecosystem. Contrary to our second hypothesis, we only found this relationship in marine and thermophilic communities. This relationship in marine communities is not especially surprising considering how many studies have observed the trade-off between the genome size and GC content of individuals in marine systems. However, our results demonstrate that these trade-offs are

detectable on a community scale and emphasizes the degree to which streamlining shapes community-averaged traits. In thermophilic communities, this relationship appeared confounded with the presence of archaea (Supplemental Fig. 2), thus making it impossible to distinguish between archaeal abundance or temperature as a driver for smaller genome size in these extreme environments. Additionally, higher temperatures might similarly result in smaller archaeal genomes (Sabath *et al.* 2013), further contributing to this signal. It is worth noting that the relationship between genome size and GC content in thermophilic communities was offset higher from marine systems, even for bacterial dominated thermophilic communities. This offset is perhaps the result of a requirement for thermal stability in hot environments which is provided by the GC triple-hydrogen bonds versus the AT double-bond (Wada and Suyama 1986; Musto *et al.* 2006).

Both GC content and average genome size in host-associated communities were low, a common feature of symbiotic bacteria (McCutcheon and Moran 2012). Although host-associated bacteria in small populations often have AT-rich genomes (Batut *et al.* 2014), the relationship between GC content and average genome size was not significant for host-associated communities. Reduced genetic flow in these communities could mean that changes in nucleotide frequency and genome size develop independently in populations. Therefore, these trends might exist within, but not between, communities. In other words, host-associated environments might produce small AT-rich genomes, but these two traits do not covary between communities as in marine systems.

Soil communities exhibited a negative relationship between average genome size and GC content. This does not necessarily exclude streamlining as a driver of genome size in soils but suggests other drivers of genome size and GC content. One explanation of this relationship is

that soil microbial communities skew towards smaller genomes with a higher GC content due to carbon limitation. A GC base pair has a carbon to nitrogen ratio of 9:8 while an AT base pair has a ratio of 10:7. A reduction in GC content therefore decreases the amount nitrogen required for DNA synthesis, which has been suggested as an explanation of the low GC content in small genomes that is commonly exhibited in marine systems, where nitrogen is often limiting (Grzymski and Dussaq 2012). In contrast, C is generally considered to be the limiting factor for growth in soil bacteria (Demoling, Figueroa and Bååth 2007; Hobbie and Hobbie 2013). A higher GC content might therefore be advantageous when C is particularly limiting. This would explain the negative correlation between genome size and GC content in soils—as smaller nutrient-limited soil bacteria would gain a stochiometric advantage from GC rich DNA. In this dataset, communities from deserts, agricultural fields, and grasslands had a smaller average genome size and higher GC content (Fig. 2). These environments tend to have lower soil and microbial carbon to nitrogen ratios than forests (Xu, Thornton and Post 2013). Similarly, bacterial communities in forests tended to have larger average genome sizes and lower GC content. Although this mechanism for nucleotide selection has not been established in soils, selection for high GC content in response to C limitation is not unfounded (Hellweger, Huang and Luo 2018; Shenhav and Zeevi 2020). Moreover, microbial communities in bare soil have been shown to have a higher GC content than in vegetated soil (Chen *et al.* 2021), and larger genomes were associated with lower GC content in a recent pangenomic study (Choudoir *et al.* 2021). It is important to note that many other environmental factors may fall along the environment gradient shown here, several of which might also influence GC content; such as temperature and moisture, which have been shown to influence nucleotide composition in terrestrial plants (Šmarda *et al.* 2014) and the genomic traits of microbes (Gravuer and Eskelinen

2017; Sorensen *et al.* 2019). Still, our data demonstrate a relationship between genomic traits in soil which is distinct to those of other systems and emphasizes the need to develop a more complete understanding of genomic features across soil microbial communities. A more thorough understanding of these relationships in soil might enhance our ability to use community-derived genomic traits in ecosystem science; for instance, in tracking growth, nutrient turnover, and microbial contributions to soil organic carbon on an ecosystem-scale. Another explanation is that fungal reads may reduce the overall GC content of a metagenome while raising estimates of average genome size. Although we attempted to avoid the influence of fungal genomes by limiting our dataset to metagenomes which were dominated by bacteria, and found that the abundance of eukaryotic reads to only slightly coincide with the relationship between average genome size (R^2=0.12) and GC content (R^2=0.14), it still is possible that even a low abundance of large fungal genomes affected our estimates. To assess this further, we applied a more stringent cut-off on the number of eukaryotic assigned reads (<1% of total) which resulted in no detectable relationship between the number or eukaryotic reads and average genome size and GC content (Supplemental Fig. 5a&b) and found that the relationship between average genome size and GC content stayed intact (Supplemental Fig. 5c).

Inconsistent with our third hypothesis, we did not find that the relative abundance of σ-factors was associated with average genome size in free-living communities. However, we did observe that marine communities maintained a lower abundance of σ-factor gene copies in comparison to other ecosystems, even when average genome size was comparable. One explanation is that the reduction of σ-factor gene copies is particularly effective in reducing reproductive costs in marine systems. Marine systems are considered to be nutrient poor relative to soils and a general reduction in the proportion of σ-factors in bacterial genomes may function as an adaptation to

nutrient constraints. We also found many trends between average genome size and the abundance of specific σ-factor genes in marine communities. In marine metagenomes, the relative abundance per genome of *rpoD* and *rpoH*, which encode for σ^D and σ^H respectively, was negatively correlated with average genome size. These trends are perhaps caused by the abundance of the streamlined SAR11 clade, which only contain σ^D and σ^H (Giovannoni 2017). Conversely, the abundance of the gene *fliA,* which encodes for the σ^{28} and regulates flagella biosynthesis (Ohnishi *et al.* 1990), increased with average genome size. This relationship reflects that found in marine systems, wherein nutrient scarcity selects for smaller, more streamlined, cells while increased nutrient availability selects for larger cells capable of chemotaxis (Lauro *et al.* 2009; Stocker 2012).

In soils, the relative abundance of many σ-factors were negatively correlated with estimates of average genome size. Most notably, we observed a decrease in the relative abundance of *rpoS* (σ^S) but no significant change in the abundance of *rpoD* (σ^D) with increasing average genome size. The balance between *rpoS* and *rpoD* may be a trade-off between stress tolerance and growth (Ferenci 2003; Nyström 2004). A higher ratio of *rpoS* to *rpoD* has been shown to increase the cell's capacity to cope with stress but limit its ability to grow on a variety of carbon sources (Ferenci 2003; King *et al.* 2004; Maharjan *et al.* 2013). We see this reflected in the environments from which the metagenomes were samples, with microbial communities from high stress environments, such as deserts, having a higher abundance or *rpoS* compared to lower-stress carbon-rich environments, such as forests (Supplemental Fig. 6).

Surprisingly, we found a high abundance of *fliA* gene copies in soil communities with smaller genomes, several of which were sourced from desert environments. Motility may be more valuable in nutrient limited environments, whereas in environments with high nutrient

inputs, nutritional competency may be more paramount. However, these results contrast with the commonly held notion that chemotaxis is most prevalent in mesic soils. One explanation is that motility may be especially important when water availability is ephemeral. A greater number of regulatory mechanisms would therefore be advantageous as it would allow for a rapid response to periodic pulses of moisture. Another possibility is that bacteria utilize biofilms surrounding fungal hyphae, or "fungal highways" (Kohlmeier *et al.* 2005), which could explain the persistence of flagellated bacteria even in xeric environments (Pion *et al.* 2013).

Finally, we found that the distribution of genomic traits estimated from soil and hot-spring communities did not follow the distribution derived from isolates—potentially due to a decoupling of traits between the individual and community level. The relationship between genome size and GC content was also substantially different between soil isolates and isolates of soil bacteria. These results indicate that certain ecosystem trade-offs may be detectable using community-derived estimates of microbial traits as opposed to isolates and showcases how relating these traits to specific environments may reveal important ecosystem-level pressures on microbial community traits.

However, it is necessary to consider that the data used for this comparison were not sourced from the same studies and the sample size was fairly limited. If genomic traits are to be used as trait-dimensions in microbial ecology, more work must be done observing the distribution of these traits both within and between communities. Further, we found that many of the studies we were able to access were collected from more specialized communities. Although we believe that the comparison of these communities still has merit in showing the range of genomic traits for particular systems, they might not accurately reflect the true distribution of these traits in their respective environments globally.

CONCLUSIONS

We found several compelling ecosystem-specific relationships between genomic traits of a microbial community, most notably with genome size, GC content and the distribution of σ-factors. Several of these relationships align with evolutionary mechanisms which relate to known drivers in these environments, such as streamlining in oceans and drift in host-associated communities. We also observed trends in soils which were not in-line with known mechanisms of genome reduction, emphasizing the need to develop an understanding of the controls of genomic features in soils. In this way our work demonstrates the importance of genomic traits in the field of microbial ecology and ecosystem science; both in their potential to assess microbial communities via ecosystem-specific trade-offs, as well as their ability to reveal new selection pressures not detectable through the analysis of individuals.

FUNDING

This work was supported by funding from the USDA National Institute of Food and Agriculture Foundational Program (award #2017-67019-26396) and additional support for PD was provided by the U.S. Department of Energy, Office of Biological and Environmental Research, Genomic Science Program LLNL 'Microbes Persist' Soil Microbiome Scientific Focus Area (award #SCW1632). Funding agencies did not play a role in study design; the collection, analysis, and interpretation of data; or writing of the manuscript.

ACKNOWLEDGEMENTS

These sequence data were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/ in collaboration with the user community. We would like to thank the following people and projects for granting us access to their data as part of this study: Jeanette Norton, Thea Whitman, Barbara Campbell, Janet Jansson, Ramunas Stepanauskas, Thomas Bianchi, Elise Morrison, Edward DeLong, William Mohn, Jonathan Raff, Robert Kelly, Nicole Dubilier, Steve Hallam, Mak Saito, David Walsh, Roland Hatzenpichler, Brett Baker, Frank Stewart, Erik Lilleskov, Devaki Bhaya, Brian Yu, Craig Cary, New Zealand Terrestrial Antarctic Biocomplexity Survey (NZTABS) supported by Antarctica New Zealand and the University of Waikato (Hamilton, New Zealand), Rick Cavicchioli, Jim Fredrickson, Jennifer Pett-Ridge, Kelly Gravuer, Emiley Eloe-Fadrosh, Charlene Kelly, Marina Kalyuzhnaya, James Tiedje, Bingbing Li, Anthony Neumann, Andreas Brune, and Gregory Dick.

We would also like to thank Megan Foley, Anita Antoninka, Carl Roybal, and Jeff Propster for their intellectual contributions to this work.

REFERENCES

Abraham BS, Caglayan D, Carrillo N V. *et al.* Shotgun metagenomic analysis of microbial communities from the Loxahatchee nature preserve in the Florida Everglades. *Environ Microbiomes* 2020;**15**:2.

Armstrong Z, Mewis K, Liu F *et al.* Metagenomics reveals functional synergy and novel polysaccharide utilization loci in the Castor canadensis fecal microbiome. *ISME J* 2018;**12**:2757–69.

Baker BJ, Lazar CS, Teske AP *et al.* Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. *Microbiome* 2015;**3**:1–12.

Bankevich A, Nurk S, Antipov D *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;**19**:455–77.

Barberán A, Ramirez KS, Leff JW *et al.* Why are some microbes more ubiquitous than others? Predicting the habitat breadth of soil bacteria. *Ecol Lett* 2014;**17**:794–802.

Bates D, Maechler M, Bolker B *et al.* Package "lme4." *dk.archive.ubuntu.com* 2020.

Battesti A, Majdalani N, Gottesman S. The RpoS-mediated general stress response in *Escherichia coli*. *Annu Rev Microbiol* 2011;**65**:189–213.

Batut B, Knibbe C, Marais G *et al.* Reductive genome evolution at both ends of the bacterial population size spectrum. *Nat Rev Microbiol* 2014;**12**:841–50.

Beam JP, Jay ZJ, Schmid MC *et al.* Ecophysiology of an uncultivated lineage of Aigarchaeota from an oxic, hot spring filamentous "streamer" community. *ISME J* 2016;**10**:210–24.

Bentley SD, Parkhill J. Comparative genomic structure of prokaryotes. *Annu Rev Genet* 2004;**38**:771–91.

Boylan SA, Redfield AR, Price CW. Transcription factor σB of *Bacillus subtilis* controls a large stationary-phase regulon. *J Bacteriol* 1993;**175**:3957–63.

Brewer TE, Handley KM, Carini P *et al.* Genome reduction in an abundant and ubiquitous soil bacterium 'Candidatus Udaeobacter copiosus.' *Nat Microbiol* 2017;**2**:16198.

Camargo AP, de Souza RSC, de Britto Costa P *et al.* Microbiomes of Velloziaceae from phosphorus-impoverished soils of the campos rupestres, a biodiversity hotspot. *Sci data* 2019;**6**:140.

Cardenas E, Kranabetter JM, Hope G *et al.* Forest harvesting reduces the soil metagenomic potential for biomass decomposition. *ISME J* 2015;**9**:2465–76.

Cardenas E, Orellana LH, Konstantinidis KT *et al.* Effects of timber harvesting on the genetic potential for carbon and nitrogen cycling in five North American forest ecozones. *Sci Rep* 2018;**8**:1–13.

Chen IMA, Chu K, Palaniappan K *et al.* IMG/M v.5.0: An integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res* 2019;**47**:D666–77.

Chen Y, Neilson JW, Kushwaha P *et al.* Life-history strategies of soil microbial communities in an arid ecosystem. *ISME J* 2021;**15**:649–57.

Choudoir MJ, Järvenpää MJ, Marttinen P *et al.* A non-adaptive demographic mechanism for genome expansion in Streptomyces. *bioRxiv* 2021, DOI: 10.1101/2021.01.09.426074.

Cobo-Simón M, Tamames J. Relating genomic characteristics to environmental preferences and ubiquity in different microbial taxa. *BMC Genomics* 2017;**18**:499.

Colatriano D, Tran PQ, Guéguen C *et al.* Genomic evidence for the degradation of terrestrial organic matter by pelagic Arctic Ocean Chloroflexi bacteria. *Commun Biol* 2018;**1**:1–9.

Demoling F, Figueroa D, Bååth E. Comparison of factors limiting bacterial growth in different soils. *Soil Biol Biochem* 2007;**39**:2485–95.

Ferenci T. What is driving the acquisition of mutS and rpoS polymorphisms in *Escherichia coli*? *Trends Microbiol* 2003;**11**:457–61.

Fernandes ND, Wu QL, Kong D *et al.* A mycobacterial extracytoplasmic sigma factor involved in survival following heat shock and oxidative stress. *J Bacteriol* 1999;**181**:4266–74.

Fierer N, Barberán A, Laughlin DC. Seeing the forest for the genes: Using metagenomics to infer the aggregated traits of microbial communities. *Front Microbiol* 2014;**5**:614.

Giovannoni SJ. SAR11 Bacteria: The Most Abundant Plankton in the Oceans. *Ann Rev Mar Sci* 2017;**9**:231–55.

Giovannoni SJ, Cameron Thrash J, Temperton B. Implications of streamlining theory for microbial ecology. *ISME J* 2014;**8**:1553–65.

Giovannoni SJ, Tripp HJ, Givan S *et al.* Genetics: Genome streamlining in a cosmopolitan oceanic bacterium. *Science (80-)* 2005;**309**:1242–5.

Gravuer K, Eskelinen A. Nutrient and rainfall additions shift phylogenetically estimated traits of soil microbial communities. *Front Microbiol* 2017;**8**:1271.

Green JL, Bohannan BJM, Whitaker RJ. Microbial biogeography: From taxonomy to traits. *Science (80-)* 2008;**320**:1039–43.

Grossman AD, Erickson JW, Gross CA. The htpR gene product of *E. coli* is a sigma factor for heat-shock promoters. *Cell* 1984;**38**:383–90.

Grzymski JJ, Dussaq AM. The significance of nitrogen cost minimization in proteomes of marine microorganisms. *ISME J* 2012;**6**:71–80.

Gweon HS, Bailey MJ, Read DS. Assessment of the bimodality in the distribution of bacterial genome sizes. *ISME J* 2017;**11**:821–4.

Hawley AK, Torres-Beltrán M, Zaikova E *et al.* A compendium of multi-omic sequence information from the Saanich Inlet water column. *Sci Data* 2017;**4**:1–11.

Hayden JD, Ades SE. The Extracytoplasmic Stress Factor, σE, Is Required to Maintain Cell Envelope Integrity in *Escherichia coli*. Sandler S (ed.). *PLoS One* 2008;**3**:e1573.

Hecker M, Schumann W, Völker U. Heat-shock and general stress response in *Bacillus subtilis*. *Mol Microbiol* 1996;**19**:417–28.

Hellweger FL, Huang Y, Luo H. Carbon limitation drives GC content evolution of a marine bacterium in an individual-based genome-scale model. *ISME J* 2018;**12**:1180–7.

Hengge R. The general stress response in gram-negative bacteria. *Bacterial Stress Responses*. Washington, DC, USA: ASM Press, 2014, 251–89.

Hershberg R, Petrov DA. Evidence that mutation is universally biased towards AT in bacteria. Nachman MW (ed.). *PLoS Genet* 2010;**6**:e1001115.

Hervé V, Liu P, Dietrich C *et al.* Phylogenomic analysis of 589 metagenome-assembled genomes encompassing all major prokaryotic lineages from the gut of higher termites. *PeerJ* 2020;**2020**:e8614.

Heurlier K, Dénervaud V, Pessi G *et al.* Negative control of quorum sensing by RpoN (σ54) in *Pseudomonas aeruginosa* PAO1. *J Bacteriol* 2003;**185**:2227–35.

Hildebrand F, Meyer A, Eyre-Walker A. Evidence of selection upon genomic GC-content in bacteria. Nachman MW (ed.). *PLoS Genet* 2010;**6**:e1001107.

Hobbie JE, Hobbie EA. Microbes in nature are limited by carbon and energy: the starving-survival lifestyle in soil and consequences for estimating microbial rates. *Front Microbiol* 2013;**4**:324.

Huete-Stauffer TM, Arandia-Gorostidi N, Alonso-Sáez L *et al.* Experimental warming decreases the average size and nucleic acid content of marine bacterial communities. *Front Microbiol* 2016;**7**:730.

Kanehisa M, Goto S. Yeast Biochemical Pathways. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30.

Karcagi I, Draskovits G, Umenhoffer K *et al.* Indispensability of horizontally transferred genes and its impact on bacterial genome streamlining. *Mol Biol Evol* 2016;**33**:1257–69.

King T, Ishihama A, Kori A *et al.* A regulatory trade-off as a source of strain variation in the species *Escherichia coli*. *J Bacteriol* 2004;**186**:5614–20.

Kirchman DL. Growth rates of microbes in the oceans. *Ann Rev Mar Sci* 2016;**8**:285–309.

Klappenbach JA, Dunbar JM, Schmidt TM. rRNA operon copy number reflects ecological strategies of bacteria. *Appl Environ Microbiol* 2000;**66**:1328–33.

Kohlmeier S, Smits THM, Ford RM *et al.* Taking the fungal highway: Mobilization of pollutant-degrading bacteria by fungi. *Environ Sci Technol* 2005;**39**:4640–6.

Konstantinidis KT, Tiedje JM. *Trends between Gene Content and Genome Size in Prokaryotic Species with Larger Genomes*., 2004.

Krause S, Le Roux X, Niklaus PA *et al.* Trait-based approaches for understanding microbial biodiversity and ecosystem functioning. *Front Microbiol* 2014;**5**:251.

Krüger K, Chafee M, Ben Francis T *et al.* In marine Bacteroidetes the bulk of glycan degradation during algae blooms is mediated by few clades using a restricted set of genes. *ISME J* 2019;**13**:2800–16.

Kuo CH, Moran NA, Ochman H. The consequences of genetic drift for bacterial genome complexity. *Genome Res* 2009;**19**:1450–4.

Kurokawa M, Seno S, Matsuda H *et al.* Correlation between genome reduction and bacterial growth. *DNA Res* 2016;**23**:517–25.

Lange R, Hengge-Aronis R. Identification of a central regulator of stationary-phase gene expression in *Escherichia coli*. *Mol Microbiol* 1991;**5**:49–59.

Lauro FM, McDougald D, Thomas T *et al.* The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci U S A* 2009;**106**:15527–33.

Lee LL, Blumer-Schuette SE, Izquierdo JA *et al.* Genus-wide assessment of lignocellulose utilization in the extremely thermophilic genus Caldicellulosiruptor by genomic, pangenomic, and metagenomic analyses. *Appl Environ Microbiol* 2018;**84**, DOI: 10.1128/AEM.02694-17.

Leung HTC, Maas KR, Wilhelm RC *et al.* Long-term effects of timber harvesting on hemicellulolytic microbial populations in coniferous forest soils. *ISME J* 2016;**10**:363–75.

Li BB, Roley SS, Duncan DS *et al.* Long-term excess nitrogen fertilizer increases sensitivity of soil microbial community to seasonal change revealed by ecological network and metagenome analyses. *Soil Biol Biochem* 2021;**160**:108349.

Li D, Liu CM, Luo R *et al.* MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015;**31**:1674–6.

Li J, Mau RL, Dijkstra P *et al.* Predictive genomic traits for bacterial growth in culture versus actual growth in soil. *ISME J* 2019:1.

Lonetto M, Gribskov M, Gross CA. The σ^{70} family: Sequence conservation and evolutionary relationships. *J Bacteriol* 1992;**174**:3843–9.

Maharjan R, Nilsson S, Sung J *et al.* The form of a trade-off determines the response to competition. van Baalen M (ed.). *Ecol Lett* 2013;**16**:1267–76.

Maresca JA, Miller KJ, Keffer JL *et al.* Distribution and diversity of rhodopsinproducing microbes in the Chesapeake Bay. *Appl Environ Microbiol* 2018;**84**, DOI: 10.1128/AEM.00137-18.

Martiny JBH, Jones SE, Lennon JT *et al.* Microbiomes in light of traits: A phylogenetic perspective. *Science (80-)* 2015;**350**, DOI: 10.1126/science.aac9323.

McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* 2012;**10**:13–26.

Mira A, Ochman H, Moran NA. Deletional bias and the evolution of bacterial genomes. *Trends Genet* 2001;**17**:589–96.

Morán XAG, Alonso-Sáez L, Nogueira E *et al.* More, smaller bacteria in response to ocean's warming? *Proc R Soc B Biol Sci* 2015;**282**:20150371.

Morris RM, Rappé MS, Connon SA *et al.* SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* 2002;**420**:806–10.

Mushinski RM, Payne ZC, Raff JD *et al.* Nitrogen cycling microbiomes are structured by plant mycorrhizal associations with consequences for nitrogen oxide fluxes in forests. 2020:1–15.

Musto H, Naya H, Zavala A *et al.* Genomic GC level, optimal growth temperature, and genome size in prokaryotes. *Biochem Biophys Res Commun* 2006;**347**:1–3.

Nayfach S, Pollard KS. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol* 2015;**16**:51.

Nayfach S, Roux S, Seshadri R *et al.* A genomic catalog of Earth's microbiomes. *Nat Biotechnol* 2020, DOI: 10.1038/s41587-020-0718-6.

Nordberg H, Cantor M, Dusheyko S *et al.* The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res* 2014;**42**, DOI: 10.1093/nar/gkt1069.

Nyström T. Growth versus maintenance: A trade-off dictated by RNA polymerase availability and sigma factor competition? *Mol Microbiol* 2004;**54**:855–62.

Ohnishi K, Kutsukake K, Suzuki H *et al.* Gene fliA encodes an alternative sigma factor specific for flagellar operons in *Salmonella typhimurium*. *MGG Mol Gen Genet* 1990;**221**:139–47.

Oksanen AJ, Blanchet FG, Kindt R *et al.* vegan: Community ecology package. 2019, DOI: 10.4135/9781412971874.n145.

Ouyang Y. Agricultural nitrogen management affects microbial communities, enzyme activities, and functional genes for nitrification and nitrogen mineralization. *All Grad Theses Diss* 2016.

Ouyang Y, Norton JM. Short-term nitrogen fertilization affects microbial community composition and nitrogen mineralization functions in an agricultural soil. *Appl Environ Microbiol* 2020;**86**, DOI: 10.1128/AEM.02278-19.

Pion M, Bshary R, Bindschedler S *et al.* Gains of bacterial flagellar motility in a fungal world. *Appl Environ Microbiol* 2013;**79**:6862–7.

Raes J, Korbel JO, Lercher MJ *et al.* Prediction of effective genome size in metagenomic samples. *Genome Biol* 2007;**8**:R10.

Raes J, Letunic I, Yamada T *et al.* Toward molecular trait□based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol Syst Biol* 2011;**7**:473.

Rinke C, Schwientek P, Sczyrba A *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 2013;**499**:431–7.

Roller BRK, Stoddard SF, Schmidt TM. Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. *Nat Microbiol* 2016;**1**:1–8.

Ronson CW, Nixon BT, Albright LM *et al. Rhizobium meliloti* ntrA (rpoN) gene is required for diverse metabolic functions. *J Bacteriol* 1987;**169**:2424–31.

Rossmassler K, Dietrich C, Thompson C *et al.* Metagenomic analysis of the microbiota in the highly compartmented hindguts of six wood- or soil-feeding higher termites. *Microbiome* 2015;**3**:56.

Sabath N, Ferrada E, Barve A *et al.* Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. *Genome Biol Evol* 2013;**5**:966–77.

Schmidt R, Gravuer K, Bossange A V *et al.* Long-term use of cover crops and no-till shift soil microbial community life strategies in agricultural soil. 2018, DOI: 10.1371/journal.pone.0192953.

Shenhav L, Zeevi D. Resource conservation manifests in the genetic code. *Science (80-)* 2020;**370**:683–7.

Šmarda P, Bureš P, Horová L *et al.* Ecological and evolutionary significance of genomic GC content diversity in monocots. *Proc Natl Acad Sci U S A* 2014;**111**:E4096-102.

Sorensen JW, Dunivin TK, Tobin TC *et al.* Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient. *Nat Microbiol* 2019;**4**:55–61.

Stocker R. Marine microbes see a sea of gradients. *Science (80-)* 2012;**338**:628–33.

Swan BK, Tupper B, Sczyrba A *et al.* Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci* 2013;**110**:11463–8.

Team RC. R: A language and environment for statistical computing. *R Found Stat Comput Vienna, Austria* 2018.

Totten PA, Cano Lara J, Lory S. The rpoN gene product of *Pseudomonas aeruginosa* is required for expression of diverse genes, including the flagellin gene. *J Bacteriol* 1990;**172**:389–96.

Vieira-Silva S, Rocha EPC. The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet* 2010;**6**:1000808.

Wada A, Suyama A. Local stability of DNA and RNA secondary structure and its relation to biological functions. *Prog Biophys Mol Biol* 1986;**47**:113–57.

Wang Q, Cen Z, Zhao J. The survival mechanisms of thermophiles at high temperatures: An angle of omics. *Physiology* 2015;**30**:97–106.

Westoby M, Gillings MR, Madin JS *et al.* Trait dimensions in bacteria and archaea compared to vascular plants. *Ecol Lett* 2021;**24**:1487–504.

Whitman T, Pepe-Ranney C, Enders A *et al.* Dynamics of microbial community composition and soil organic carbon mineralization in soil following addition of pyrogenic and fresh organic matter. *ISME J* 2016;**10**:2918–30.

Wilhelm RC, Cardenas E, Leung H *et al.* Data Descriptor: A metagenomic survey of forest soil microbial communities more than a decade after timber harvesting Background & Summary. 2017a, DOI: 10.1038/sdata.2017.92.

Wilhelm RC, Cardenas E, Leung H *et al.* Long-term enrichment of stress-tolerant cellulolytic soil populations following timber harvesting evidenced by multi-omic stable isotope probing. *Front Microbiol* 2017b;**8**:537.

Wilhelm RC, Cardenas E, Maas KR *et al.* Biogeography and organic matter removal shape long-term effects of timber harvesting on forest soil microbial communities. *ISME J* 2017c;**11**:2552–68.

Williams TJ, Allen MA, Berengut JF *et al.* Shedding Light on Microbial "Dark Matter": Insights Into Novel Cloacimonadota and Omnitrophota From an Antarctic Lake. *Front Microbiol* 2021;**12**:2947.

Xu X, Thornton PE, Post WM. A global analysis of soil microbial biomass carbon, nitrogen and phosphorus in terrestrial ecosystems. *Glob Ecol Biogeogr* 2013;**22**:737–49.

Yooseph S, Nealson KH, Rusch DB *et al.* Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature* 2010;**468**:60–6.

Zuber U, Drzewiecki K, Hecker M. Putative sigma factor sigI (ykoZ) of *Bacillus subtilis* is induced by heat shock. *J Bacteriol* 2001;**183**:1472–5.

LIST OF FIGURES

Figure 1:

Average genome size and GC-content calculated from environmental metagenomes. (**A**) Boxplots of the average genome size (Mbp) of microbial communities in different ecosystems. (**B**) Boxplots showing GC-% between systems. **(C)** GC-% as a function of average genome size (Mbp) of a metagenome, separated by system. Point shape and outline represent source system; point fill represents system including thermophilic samples with archaea.

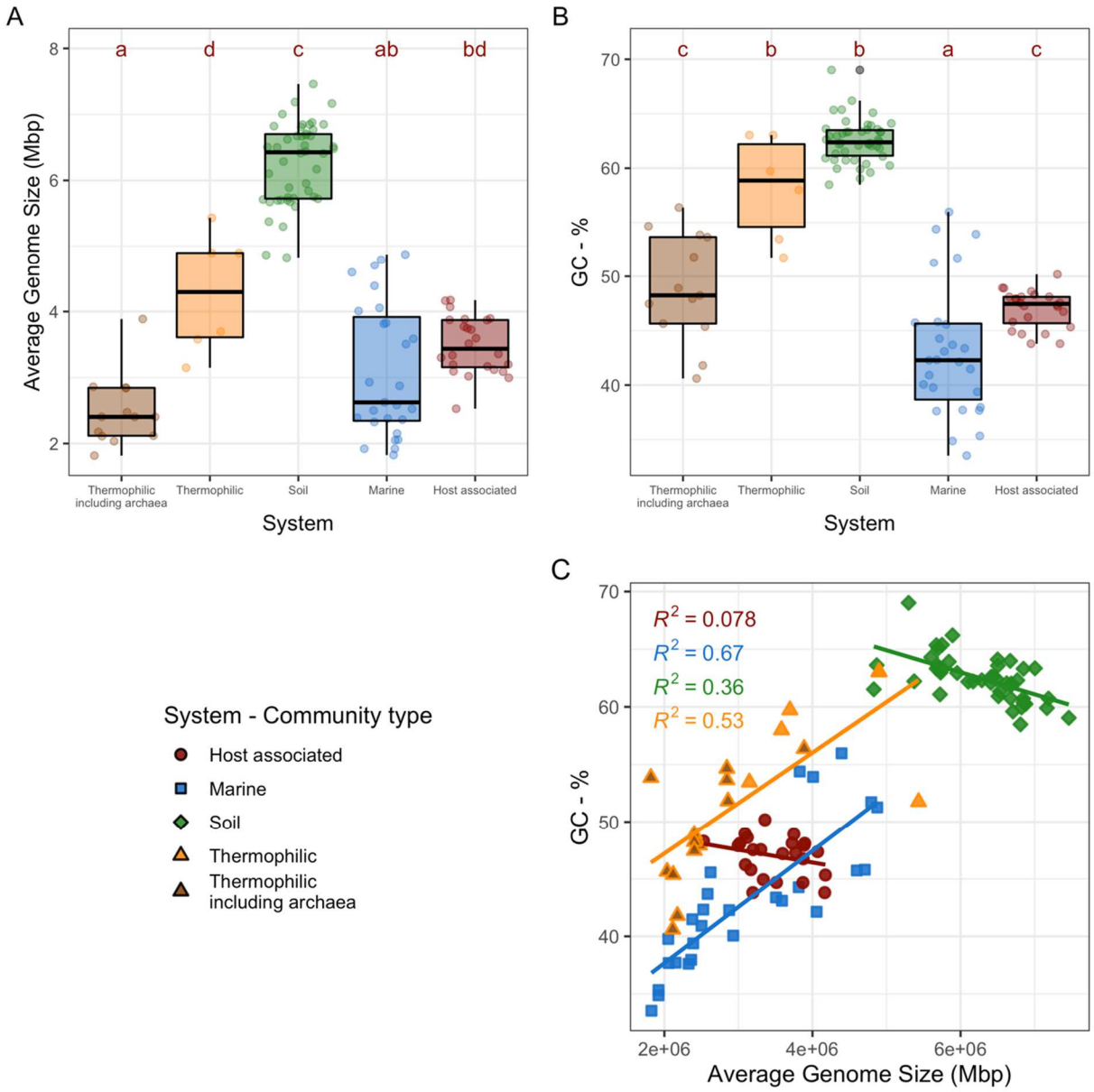

A
a
d
c
ab
bd
Average Genome Size (Mbp)
8
6
4
2
Thermophilic including archaea
Thermophilic
Soil
Marine
Host associated
System
B
c
b
b
a
c
GC - %
70
60
50
40
Thermophilic including archaea
Thermophilic
Soil
Marine
Host associated
System
C
R^2 = 0.078
R^2 = 0.67
R^2 = 0.36
R^2 = 0.53
GC - %
70
60
50
40
2e+06
4e+06
6e+06
Average Genome Size (Mbp)
System - Community type
Host associated
Marine
Soil
Thermophilic
Thermophilic including archaea

Figure 2:

GC content (%) as a function of average genome size (Mbp) in soils, with color indicating source environment.

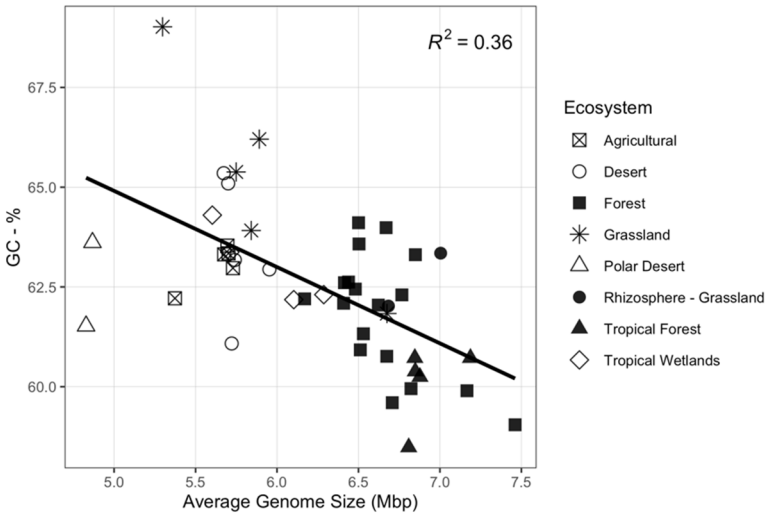


Figure 3:

The relationship and distribution of genome size and GC content for isolates and metagenomic averages for each system. In each panel, metagenomes (dark circles) are plotted against bacterial (light squares) and archaeal (light triangles) isolates. Regression lines between genome size and GC-% are shown for both metagenomes (dark lines) and isolates (light lines). Marginal density plots show the distributions of GC-% (right) and genome size (top) for isolates (light) and metagenomic averages (dark).

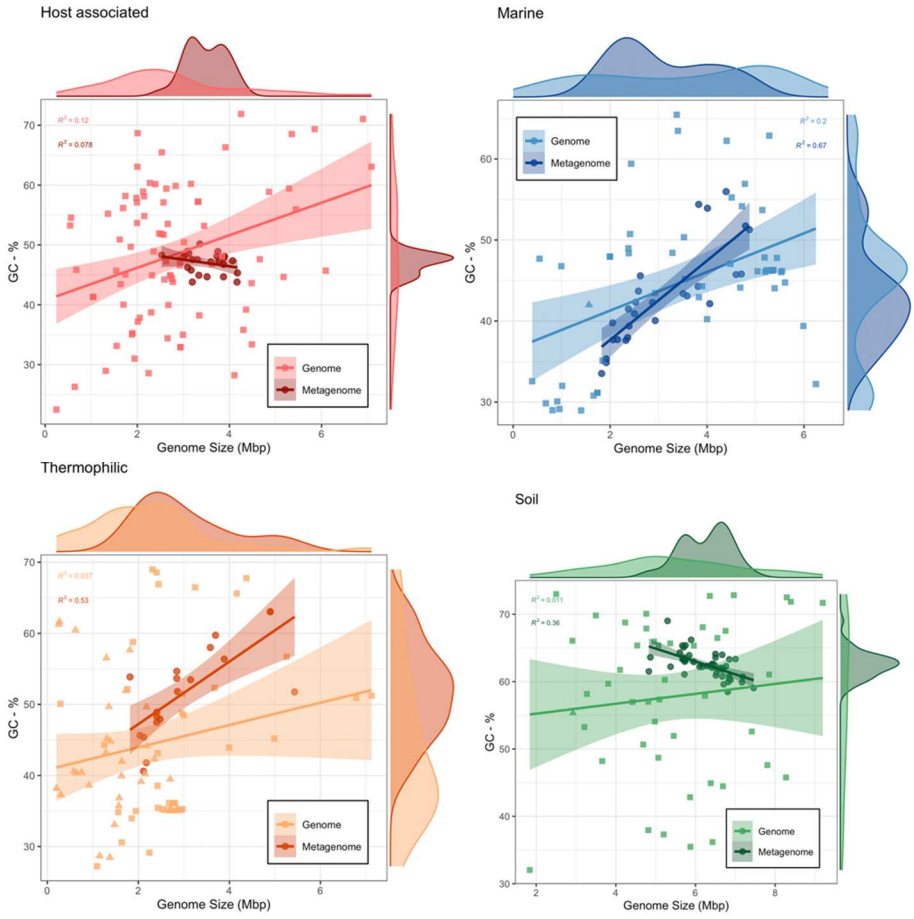


Figure 4:

The relative abundance σ-factors in a metagenome separated by ecosystem. Each bar represents the abundance of σ-factors in a single metagenome, and metagenomes are ordered from smallest to largest (left to right) for each ecosystem.

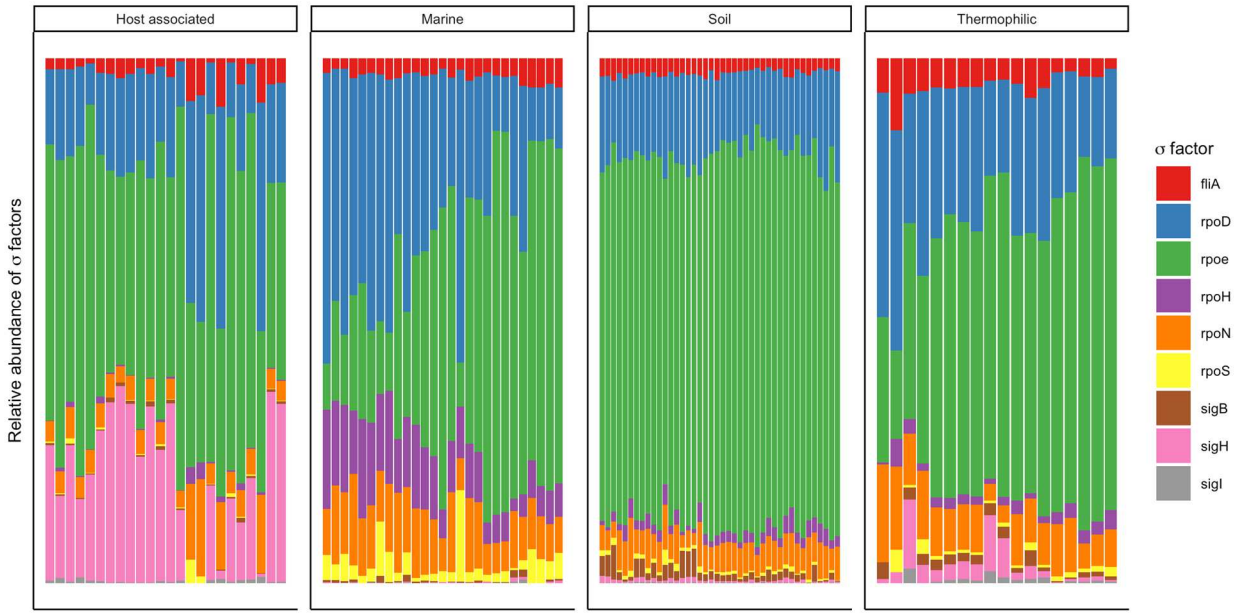
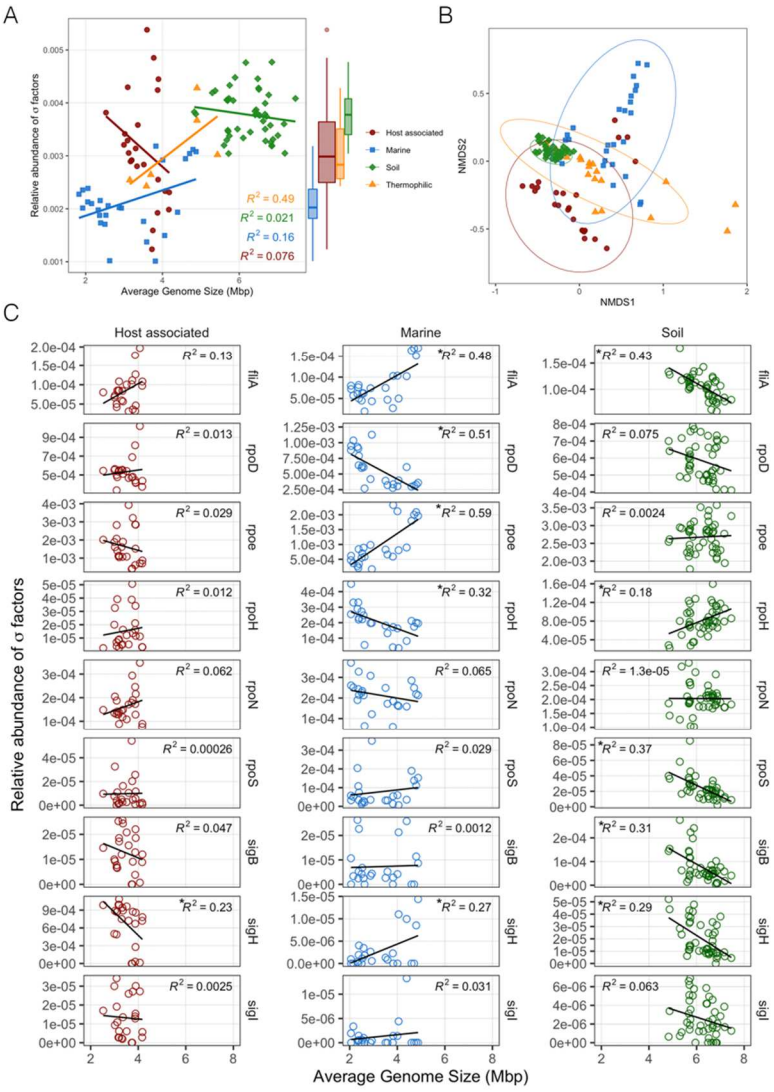


Figure 5:

The relative abundance of σ-factors (σ-factor count / gene count) as a function of average genome size and system. **(A)** The relative abundance of all σ-factors (total σ-factor count / gene count) in a metagenome against average genome size. Source environment indicated by color for host associated (red), soil (green), thermophilic (orange) and marine (blue) communities. **(B)** NMDS of Bray-Curtis distance of the relative abundance of σ-factors (σ-factor count / total gene count) from a metagenome. **(C)** The relative abundance (σ-factor count / total gene count) of 9 σ-factors (rows) versus average genome size, separated by environment (columns). Statistical significance of a relationship ($p < 0.05$) is indicated with an asterisk.



LIST OF TABLES

Gene name, description, and KEGG ortholog identifier (K numbers) for each σ-factor used in the analysis.

***σ*-factor gene**	**Functions regulated by *σ*-factor**	**K Number**
rpoD	Primary sigma factor, "Housekeeping" (Lonetto, Gribskov and Gross 1992)	KO:K03086
rpoE	Envelope stress (Hayden and Ades 2008)	KO:K03088
fliA	Flagella biosynthesis (Ohnishi *et al.* 1990)	KO:K02405
rpoH	Heat shock (Grossman, Erickson and Gross 1984)	KO:K03089
sigI	Heat shock (Zuber, Drzewiecki and Hecker 2001)	KO:K03093
sigH	Heat shock, oxidative stress (Fernandes *et al.* 1999)	KO:K03091
rpoN	Nitrogen assimilation (Ronson *et al.* 1987; Totten, Cano Lara and Lory 1990), Motility (Totten, Cano Lara and Lory 1990), Quorum sensing (Heurlier *et al.* 2003)	KO:K03092
rpoS	Stress response (Battesti, Majdalani and Gottesman 2011; Hengge 2014), Stationary phase (Lange and Hengge-Aronis 1991)	KO:K03087
sigB	Stress response (Hecker, Schumann and Völker 1996) Stationary phase (Boylan, Redfield and Price 1993)	KO:K03090

CHAPTER 4

EDAPHIC CONTROLS ON GENOME SIZE AND GC CONTENT OF BACTERIA IN SOIL MICROBIAL COMMUNITIES

AUTHORS:

Peter F. Chuckran[1]

Cody Flagg[2]

Jeffrey Propster[1]

William A. Rutherford[3]

Ella Sieradzki[4]

Steven J. Blazewicz[5]

Bruce Hungate[1]

Jennifer Pett-Ridge[5,6]

Egbert Schwartz[1]

Paul Dijkstra[1]

[1] Center for Ecosystem Science and Society (ECOSS) and Department of Biological Sciences, Northern Arizona University, Flagstaff, Arizona, USA

[2] National Ecological Observation Network (NEON), Boulder, Colorado, USA

[3] School of Natural Resources and the Environment, University of Arizona, Tucson, Arizona, USA

[4] Department of Environmental Science, Policy, and Management, University of California, Berkeley, California, USA

[5] Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, California, USA

[6] Life & Environmental Sciences Department, University of California Merced, Merced, California, USA

Competing interest statement:

The authors have no competing interests to disclose.

ABSTRACT

Genomic traits, such as genome size, GC content, codon usage, and amino acid content, shed insight into the evolutionary processes of bacteria and selective forces behind microbial community composition. Nutrient limitation has been shown to reduce bacterial genome size and influence nucleotide composition, yet little research has been conducted in the soil environment, and the factors which shape soil bacterial genomic traits remain largely unknown. Here we determined average genome size, GC content, codon usage, and amino acid content from 398 soil metagenomes across the United States from the National Ecological Observation Network (NEON) and observed the distribution of these traits across numerous environmental gradients. We found that genomic trait averages were most strongly related to pH, which we suggest results in both physiological constraints on growth as well as affects availability of nutrients in soil. Low pH soils had higher carbon to nitrogen ratios (C:N) and tended to have communities with larger genomes and lower GC-content, potentially a result of increased physiological stress and increased metabolic diversity. Conversely, smaller genomes with high GC content were associated with high pH and low soil carbon to nitrogen ratios, indicating potential resource driven selection against carbon-rich AT base pairs. We found that this relationship of nutrient conservation also applied to amino acid stoichiometry, where bacteria in soils with C:N ratios tended to code for amino acids with lower C:N. Together, these relationships point towards fundamental mechanisms which underpin nucleotide and amino acid selection in soil bacterial communities.

MAIN TEXT

In bacteria, nutrient constraints exert influence on traits such as genome size, GC content, codon frequency, and amino acid content [1–3]. For free living bacteria, low nutrient concentrations often select for genomic traits which reduce the cost of reproduction, such as low GC content and smaller genomes [3]. Since the AT base pair has a carbon to nitrogen ratio (C:N) of 10:7 (1.42), whereas a GC base pair has a C:N of 9:8 (1.13), the AT base pair is more advantageous in nitrogen-limited environments. However, much of the existing and foundational literature on processes controlling genomic traits in free-living bacteria are based on the study of marine isolates [4] and aspects of this framework may not cleanly transpose onto soil bacteria. For example, we had previously found that community-averaged GC content and genome size were positively correlated between marine metagenomes, but surprisingly negatively correlated between metagenomes collected from soils [5]. Since the growth of soil bacteria is thought to be more limited by carbon than nitrogen [6, 7], we hypothesized that the distribution of genomic traits in soil bacteria might exhibit unique patterns reflecting carbon limitation: specifically, higher GC content and smaller genomes when carbon availability is low.

To better understand the relationship between genomic traits of soil bacteria and edaphic characteristics, we analyzed 398 metagenomes collected and sequenced by the US-based National Ecological Observation Network (NEON) [8] across a broad geographic scale (Fig. 1A) and analyzed genomic traits alongside a range of environmental and soil properties. For each metagenome, QC filtered reads were assembled into contigs and annotated. Bacterial contigs were then used to generate a community-level estimate of bacterial GC content, codon frequency, and amino acid content. Using the chemical formulas for each amino acid, we calculated the total relative carbon and nitrogen content of amino acids for each metagenome

(see Methods). Average genome size of a community was estimated from the number of single copy genes recovered from the QC filtered metagenomic reads. Since this estimate uses all QC filtered reads and may therefore be biased by the presence of large fungal genomes, we calculated an additional estimate of bacterial genome size using 16S rRNA gene datasets produced from the same soil samples (sequenced separately from the metagenomes). By aligning community taxonomic data against a database of genome sizes from high-quality isolates, we were able to calculate a community weighted mean of genome size for the bacteria in a community. Genomic traits were then paired with NEON environmental data (soil chemistry, climate, etc.), as well as select meteorological data accessed from GRIDMET [9].

We hypothesized that microbial communities in carbon limited environments would exhibit smaller genome sizes, higher GC content, and amino acid composition with a lower carbon to nitrogen ratio (C:N). To test this, we assessed the relationship between extractable soil C:N (C_{extr}:N_{extr}) and genomic traits across all NEON sites (Fig. 1A). Most of the communities with small genome size and high GC content were located in the mountain-west. We further found a negative correlation between GC content and C_{extr}:N_{extr} ($p < 0.001$; Fig. 1B), and a positive correlation between C_{extr}:N_{extr} and both estimates of average genome size ($p < 0.001$; Fig. 1C). The C:N of the sum of all coded amino acids was positively correlated with soil C_{extr}:N_{extr} (R^2=0.24, $p < 0.001$, Fig. 1D), and closely tracked metagenome GC content (R^2=0.51, $p < 0.001$, Fig. 1E)—reflecting the resource alignment between the stoichiometry of nucleic acids in codons and their corresponding amino acids [10]. We found that synonymous codon usage skewed towards codons with a higher GC content—perhaps best represented by the strong preference for guanine and cytosine at fourfold degenerate sites ($p < 0.01$; Fig. 1F). Preferential selection for codons with higher AT content was most pronounced where soil C:N was high. For each amino

acid, codons with higher GC were more often negatively correlated with soil C_{extr}:N_{extr} compared to codons with lower GC, which more often were positively correlated with soil C_{extr}:N_{extr} (Fig. 1G&H).

These results are in line with our original hypothesis wherein lower C:N lead to communities averaging smaller genomes, higher GC, and lower C:N of amino acids. However, genomic traits in soil microbial communities may also be driven by other environmental factors, such as temperature [11, 12] and pH [13]. To assess the relationships between genomic traits and other environmental drivers, we used a machine learning, random-forest model approach to determine the environmental variables which explain the most variance in GC content and average genome size. With this model, we assessed the importance of over 100 environmental factors and geographic range in shaping genomic features.

Random forest models indicated that GC content and the average genome size of a community were most strongly related to soil pH (Fig. 2A), where soils with low pH fostered communities with low GC content (Fig. 2B) and larger average genome size (Fig. 2C). Although changes in the average genome size of a community as determined by single copy genes could be biased by large fungal genomes at low pH, we additionally found a relationship between pH and genome size when genome size was determined by aligning 16S gene sequences with a database of isolates of known size.

Soil pH represents the intersection of numerous environmental vectors and, accordingly, we hypothesize that there are several mechanisms underpinning the relationship between pH and genomic traits. First, low pH causes physiological stress in soil bacteria and is often associated with a greater number of repair mechanisms, such as chaperones [14]. This might preferentially select for bacteria with a greater investment in stress alleviation and maintenance, and thus larger

genomes. Second, low pH is often associated with the accumulation of soil organic carbon (SOC). Since soil pH is largely driven by the balance between precipitation and evapotranspiration [15], low pH often coincides with greater precipitation excess and primary production. Higher biomass inputs into acidic soils, combined with a reduction in the decomposition rate due to low pH, results in the build-up of SOC. The accumulation of SOC not only alleviates carbon limitation—which may reduce GC content—but also potentially favors larger genomes with increased metabolic diversity. It has been suggested that the requirement for increased metabolic diversity might explain why soil bacterial genomes tend to have large genomes [16] and, similarly, we found that soils with lower pH were associated with higher C_{extr}:N_{extr} (Fig. 2D), as well as larger genomes and lower GC content (Fig. 2D). Third, genomic traits in soil bacteria may relate to other forms of stress coinciding with pH. Aridity has been shown to drive streamlining in certain soil bacteria [17] and, as discussed above, influences the pH in soil. We did find a positive relationship between mean annual precipitation and average genomes size (Supplemental Fig. 1), although the relationship was not as strong as that for pH or soil C_{extr}:N_{extr}. Previous work has shown relationships between precipitation, pH, and genome size [13], and in a previous analysis we found that soil metagenomes collected from both hot and cold deserts often had smaller genomes and greater GC content than soils collected in more mesic systems [5].

Our work demonstrates that soil pH determines the broad-scale distribution of genomic traits between soil bacterial communities, which we suggest can be attributed to pH being a metric which captures multiple environmental parameters, such as soil nutrients, precipitation patterns, and physiological stress. Soil pH is well established driver of community composition in soils [18] and our results emphasize the degree to which pH dictates belowground microbial

life and perhaps influence evolution of soil bacteria. Additionally, we found several trends which suggest that selection pressure in soil bacterial communities might reflect carbon limitation. Whereas marine communities tend to harbor small AT-rich bacteria in response to nitrogen limitation [3], we found that soils low in carbon tended to select for communities where genomes were smaller but high in GC content. These results are derived from community averages and more work must be done to uncover both the mechanisms and taxonomic level at which such changes in genomic traits occur. However, it is evident that the distribution of genomic traits in soil remains distinct from marine systems, and that physiological stress and nutrient demand are likely written into the DNA of soil bacteria.

METHODS

Data for this project was gathered from the National Ecological Observation Network (NEON)—an observational network collecting ecological data from across the United States, funded by the US National Science Foundation. NEON maintains 81 field sites in number of distinct biomes across the US. Terrestrial sites include a central meteorological station as well as many dispersed plots from which samples are collected [19]. A full description of NEON sites can be found at https://www.neonscience.org/field-sites/about-field-sites.

Metagenomic traits

From the NEON data portal (https://www.neonscience.org/data), we accessed the metadata for all available metagenomes as of January 2021. From these data we selected 398 metagenomes which demonstrated the greatest read depth while maximizing the number of collection sites (43 total). In January of 2021, selected metagenomes were downloaded from the NEON data portal to the high-performance computing cluster at Northern Arizona University.

Raw reads were QC filtered using Trimmomatic v0.39 (parameters: TruSeq2-PE.fa:2:30:10 LEADING:10 TRAILING:10 SLIDINGWINDOW:10:20 MINLEN:50) [20]. We then used the program MicrobeCensus [21] to determine the average genome size of a microbial community from the QC-filtered reads. MicrobeCensus estimates the total number of single copy genes to create an estimate of average genome size from unassembled reads. Contigs were assembled from QC-filtered reads using MEGAHIT v1.2.9 (parameters: --k-list 21,29,39,59,79 --min-contig-len 400) [22] and read depth for each contig was determined using BBMap v38.87 [23]. Open reading frames (ORFs) were then identified using Prodigal v2.6.3 (parameters: -p meta) [24] and taxonomy was assigned to each ORF using Kaiju v1.7.4 using the default parameters [25].

Using the read depth for each contig and the estimated taxonomic identity for each ORF, we calculated a depth adjusted GC content for the bacterial reads in a metagenome. Similarly, we calculated the depth-adjusted amino acid content for each metagenome. Using known chemical formulas for each amino acid and the depth-adjusted amino acid content, we calculated an estimate of the total C:N of the amino acids in each metagenome. For amino acids with fourfold degenerative sites (alanine, glycine, proline, threonine, valine), we calculated the frequency of nucleotides at the third position and averaged these values across the selected amino acids. This provided a metagenome-level estimate of nucleotide frequencies for fourfold degenerative sites.

Since our estimate of average genome size for a metagenome was based on the QC filtered reads, large fungal genomes could potentially bias our estimate. To assure that trends we observed were not solely driven by the presence of large fungal genomes, we also estimated average genome size using a community-weighted mean derived from aligning bacterial 16S rRNA gene sequences against a database of genomes of known size. In August 2021 we

downloaded all available 16S gene sequences and associated metadata from the Genome Taxonomy Database [26] and compiled a database consisting of high quality (>95% complete, <5% redundant) bacterial genomes. We then accessed 16S rRNA gene datasets from the NEON data portal derived from the same plots as our metagenomes. Operational taxonomic units (OTUs; clustered at 97%) identified by NEON were then aligned to our high-quality database using BLAST. The best alignment (at >99% identity) was used as the assigned taxonomy. The genome size of each assigned OTU was then adjusted for read depth and used to calculate a community-weighted mean genome size with mean coverage (read depth) as the weight factor.

Analysis with environmental characteristics

From the NEON data portal and meteorological data, we gathered 205 environmental parameters associated with each site and date of collection. A full list of the NEON data products used in this analysis can be found in Supplemental Table 1. Data were accessed between January and April 2021. Where possible, edaphic characteristics were paired with metagenomes from the same soil sample. Otherwise, metagenomes were paired with data from plot or site level averages. Microbial PLFA biomarkers were used to calculate fungal to bacterial biomass ratios [27, 28]. Standard precipitation and standard precipitation evapotranspiration indices (SPI and SPEI) were gathered from GRIDMET [9] using Google Earth Engine [29]. Regression analysis was used to determine the relationship between soil carbon and genomic traits. All models were constructed in R (v 3.6.1) [30].

Random forest models were constructed using the Tidymodels [31] and ranger [32] packages in R, with the objective of determining the environmental parameters which most strongly influenced genomic traits. A non-parametric, machine learning/random forest regression

model approach was selected given its proven ability to handle non-linear, complex interactions in space and time across multiple predictor variables [33]. Prior to tuning the random forest hyperparameters, predictors which did not provide data for at least two-thirds of metagenomes were dropped, and the remaining variables were scaled and centered prior to analysis. The full dataset contained several environmental measurements which were either very similar or functionally identical metrics collected at multiple levels (e.g., site, plot, and soil core). Redundant predictors were removed by first running a random forest which included all available predictors, which were then ranked by variable importance. We calculated Pearson's correlation coefficients between all predictors and then removed those which were highly correlated ($R^2 > 0.8$) with a predictor which explained more variability. We then ran a random forest model which included only remaining predictors.

Upon publication, all code and corresponding data generated through this analysis will be made publicly available through Github.

FUNDING

This work was supported by funding from the USDA National Institute of Food and Agriculture Foundational Program (award #2017-67019-26396). Support for SB, ES, JP, PD, and BH was provided by the U.S. Department of Energy, Office of Biological and Environmental Research, Genomic Science Program LLNL 'Microbes Persist' Soil Microbiome Scientific Focus Area (award #SCW1632). Work conducted at LLNL was conducted under the auspices of the US Department of Energy under Contract DE-AC52-07NA27344. Funding agencies did not play a role in study design; the collection, analysis, and interpretation of data; or writing of the manuscript.

ACKNOWLEDGEMENTS:

We would like to thank Anita Antoninka, Michaela Hayer, Alicia Purcell, Junhui Li, Megan Foley, Raina Fitzpatrick, Bram Stone, Victoria Monsaint-Queeney, and Carl Roybal for their intellectual contributions to this work.

REFERENCES

1. Shenhav L, Zeevi D. Resource conservation manifests in the genetic code. *Science (80-)* 2020; **370**: 683–687.

2. Batut B, Knibbe C, Marais G, Daubin V. Reductive genome evolution at both ends of the bacterial population size spectrum. *Nat Rev Microbiol* 2014; **12**: 841–850.

3. Giovannoni SJ, Cameron Thrash J, Temperton B. Implications of streamlining theory for microbial ecology. *ISME J* 2014; **8**: 1553–1565.

4. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, et al. Genetics: Genome streamlining in a cosmopolitan oceanic bacterium. *Science (80-)* 2005; **309**: 1242–1245.

5. Chuckran PF, Hungate B, Schwartz E, Dijkstra P. Soil, ocean, hot spring, and host-associated environments reveal unique selection 1 pressures on genomic features of bacteria in microbial communities 2 3 AUTHORS: 4. *bioRxiv* 2021; 2021.04.05.438506.

6. Soong JL, Fuchslueger L, Marañon-Jimenez S, Torn MS, Janssens IA, Penuelas J, et al. Microbial carbon limitation: The need for integrating microorganisms into our understanding of ecosystem carbon cycling. *Glob Chang Biol* 2020; **26**: 1953–1961.

7. Hobbie JE, Hobbie EA. Microbes in nature are limited by carbon and energy: the starving-survival lifestyle in soil and consequences for estimating microbial rates. *Front Microbiol* 2013; **4**: 324.

8. National Ecological Observatory Network (NEON). Soil microbe metagenome sequences (DP1.10107.001). 2021.

9. Abatzoglou JT. Development of gridded surface meteorological data for ecological applications and modelling. *Int J Climatol* 2013; **33**: 121–131.

10. Bragg JG, Hyder CL. Nitrogen versus carbon use in prokaryotic genomes and proteomes. *Proc R Soc B Biol Sci* 2004; **271**: 374–377.

11. Sabath N, Ferrada E, Barve A, Wagner A. Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. *Genome Biol Evol* 2013; **5**: 966–977.

12. Sorensen JW, Dunivin TK, Tobin TC, Shade A. Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient. *Nat Microbiol* 2019; **4**: 55–61.

13. Gravuer K, Eskelinen A. Nutrient and rainfall additions shift phylogenetically estimated traits of soil microbial communities. *Front Microbiol* 2017; **8**: 1271.

14. Malik AA, Puissant J, Buckeridge KM, Goodall T, Jehmlich N, Chowdhury S, et al. Land use driven change in soil pH affects microbial carbon cycling processes.

15. Slessarev EW, Lin Y, Bingham NL, Johnson JE, Dai Y, Schimel JP, et al. Water balance creates a threshold in soil pH at the global scale. *Nature* 2016; **540**: 567–569.

16. Barberán A, Ramirez KS, Leff JW, Bradford MA, Wall DH, Fierer N. Why are some microbes more ubiquitous than others? Predicting the habitat breadth of soil bacteria. *Ecol Lett* 2014; **17**: 794–802.

17. Simonsen AK. Environmental stress leads to genome streamlining in a widely distributed species of soil bacteria. *ISME J 2021* 2021; 1–12.

18. Fierer N. Embracing the unknown: Disentangling the complexities of the soil microbiome. *Nat Rev Microbiol* 2017; **15**: 579–590.

19. Hinckley ELS, Bonan GB, Bowen GJ, Colman BP, Duffy PA, Goodale CL, et al. The soil and plant biogeochemistry sampling design for The National Ecological Observatory Network. *Ecosphere* 2016; **7**: e01234.

20. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 2014; **30**: 2114–2120.

21. Nayfach S, Pollard KS. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol* 2015; **16**: 51.

22. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015; **31**: 1674–1676.

23. Bushnell B. BBTools Software Package. https://sourceforge.net/projects/bbmap/. .

24. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010; **11**: 1–11.

25. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. 2016.

26. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A, Hugenholtz P. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* 2021.

27. Willers C, Jansen van Rensburg PJ, Claassens S. Phospholipid fatty acid profiling of microbial communities-a review of interpretations and recent applications. *J Appl*

Microbiol 2015; **119**: 1207–1218.

28. Orwin KH, Dickie IA, Holdaway R, Wood JR. A comparison of the ability of PLFA and 16S rRNA gene metabarcoding to resolve soil community change and predict ecosystem functions. *Soil Biol Biochem* 2018; **117**: 27–35.

29. Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens Environ* 2017; **202**: 18–27.

30. Team RC. R: A language and environment for statistical computing. *R Found Stat Comput Vienna, Austria* 2018.

31. Kuhn M, Wickham H. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. 2020.

32. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J Stat Softw* 2015; **77**.

33. Breiman L. Random forests. *Mach Learn* 2001; **45**: 5–32.

LIST OF FIGURES

Figure 1:

Distribution of genomic traits across sites and soil extractable carbon and extractable nitrogen ratios (C_{extr}:N_{extr}); **(A)** Geographic distribution of sites, with mean bacterial GC-% and estimated average genome size. **(B)** Relationship between C_{extr}:N_{extr} and bacterial GC content (%). **(C)** Relationship between C_{extr}:N_{extr} and genome size, estimated from the number of single copy genes per metagenome (black circles) and from 16S rRNA gene datasets and genome size estimated from isolates (open circles)**. (D)** Relationship between C_{extr}:N_{extr} and bacterial amino acid C:N ratios averaged per metagenome. **(E)** Relationship between bacterial GC-% and pooled amino acid C:N ratios**. (F)** The distribution of nucleotides at the third position in fourfold degenerate codons across all metagenomes, with letters corresponding to groups identified via Tukey's post-hoc test. **(G)** The relationship between codon frequency and soil C_{extr}:N_{extr} shown for aspartic acid and glutamic acid, with number of GC base pairs in each codon being indicated by color. **(H)** The relationship between codon frequency and soil C_{extr}:N_{extr} shown for each codon, with color indicating the slope of the relationship and an asterisk indicating significance ($p < 0.05$). Codons are arranged left to right in increasing number of GC base pairs.

Figure 1:

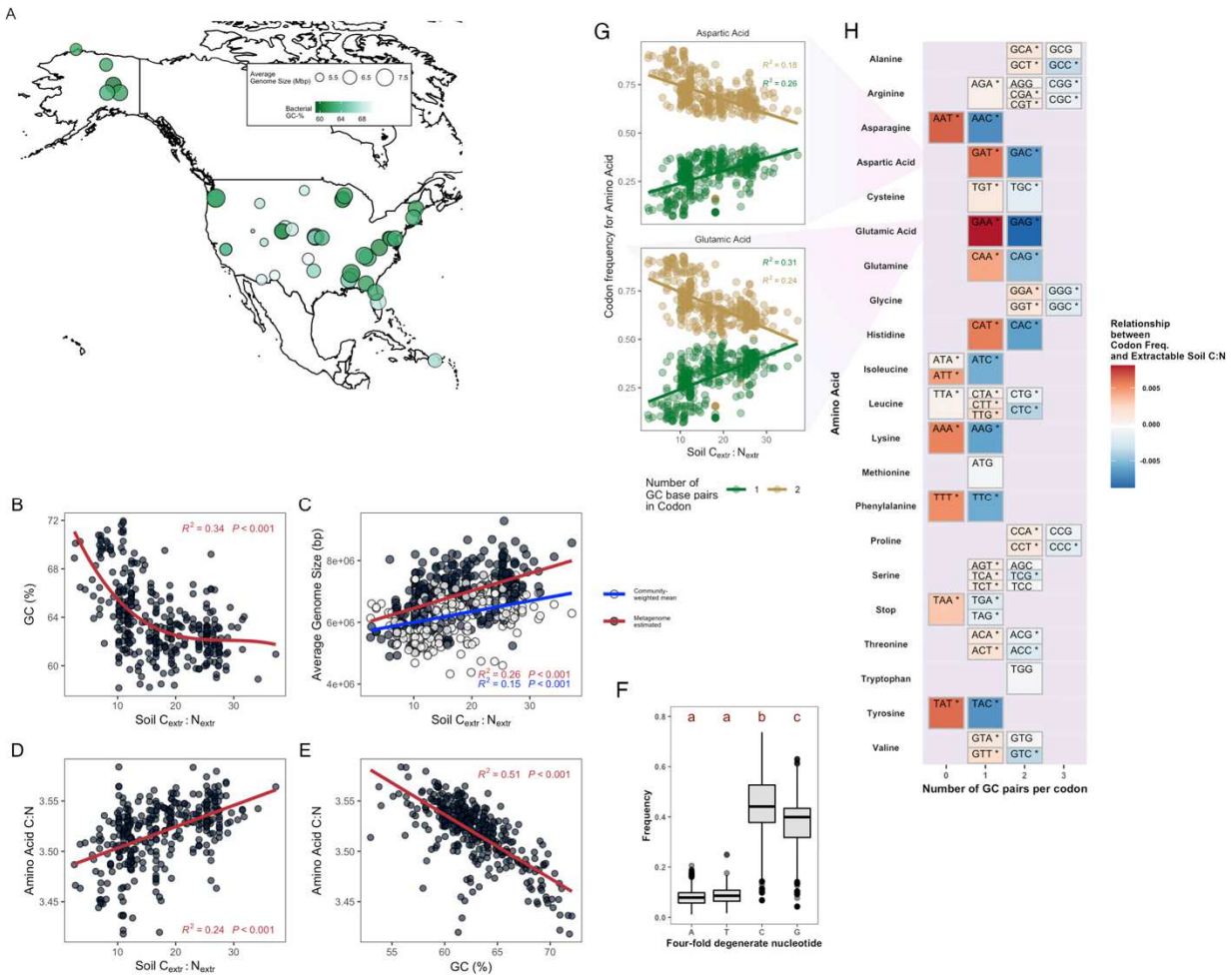
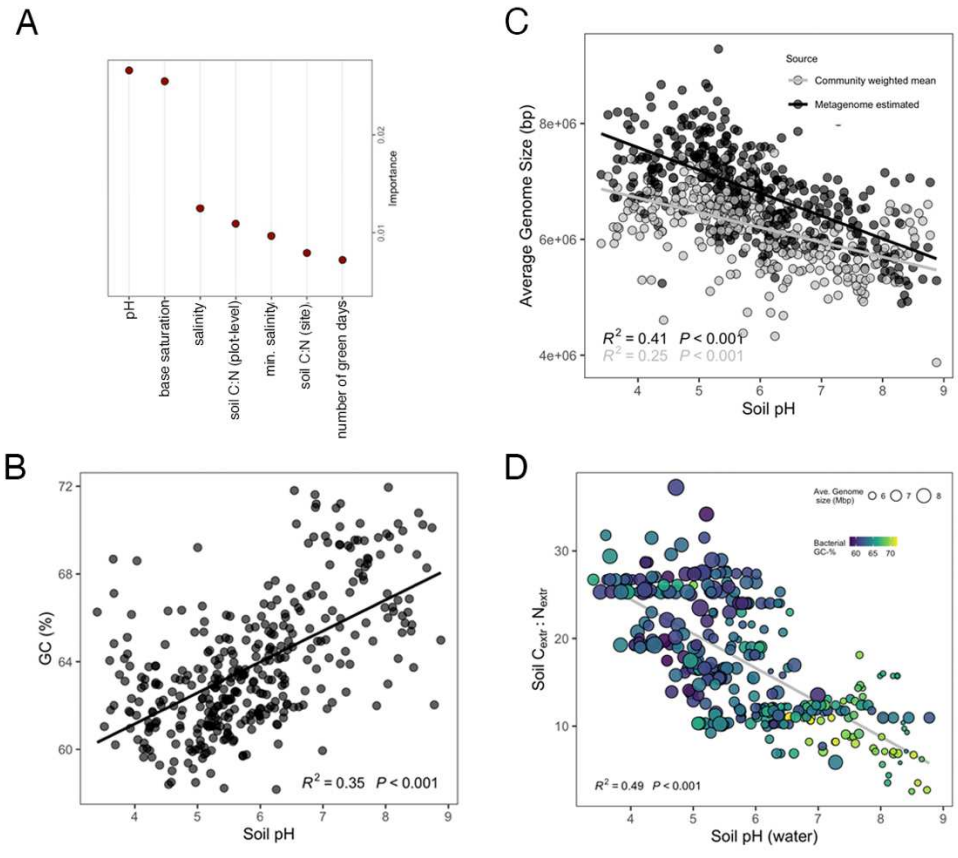

A
Average Genome Size (Mbp)
5.5
6.5
7.5
Bacterial GC-%
60
64
68
B
R^2 = 0.34 P < 0.001
GC (%)
Soil C_{extr} : N_{extr}
C
Average Genome Size (bp)
R^2 = 0.26 P < 0.001
R^2 = 0.15 P < 0.001
Soil C_{extr} : N_{extr}
D
Amino Acid C:N
R^2 = 0.24 P < 0.001
Soil C_{extr} : N_{extr}
E
Amino Acid C:N
R^2 = 0.51 P < 0.001
GC (%)
G
Aspartic Acid
R^2 = 0.18
R^2 = 0.26
Glutamic Acid
R^2 = 0.31
R^2 = 0.24
Codon frequency for Amino Acid
Soil C_{extr} : N_{extr}
Number of GC base pairs in Codon
1
2
Community-weighted mean
Metagenome estimated
F
a
a
b
c
Frequency
A
T
C
G
Four-fold degenerate nucleotide
H
Alanine
Arginine
Asparagine
Aspartic Acid
Cysteine
Glutamic Acid
Glutamine
Glycine
Histidine
Isoleucine
Leucine
Lysine
Methionine
Phenylalanine
Proline
Serine
Stop
Threonine
Tryptophan
Tyrosine
Valine
Amino Acid
Number of GC pairs per codon
Relationship between Codon Freq. and Extractable Soil C:N
0.005
0.000
-0.005

Figure 2:

Results from the random forest model and relationships between soil pH and genomic traits; **(A)** Variable importance plot for the top 10 environmental parameters predicting GC content from the random forest model (RMSE = 0.017; R^2 = 0.66). **(B)** The relationship between soil pH and GC content (%) of bacterial contigs. **(C)** The relationship between soil pH and both average genome size (derived from metagenomes) and community-weighted genome size (derived from the identification of 16S rRNA genes). **(D)** The relationship between soil pH and soil C_{extr}:N_{extr} with points colored by bacterial GC content and point size corresponding to average genome size.



CHAPTER 5
CODON OPTIMIZATION IN SOIL METATRANSCRIPTOMES IN RESPONSE TO CARBON INPUTS AND STRESS

ABSTRACT

Codon and nucleotide frequencies are known to influence the rate of transcription for functional genes, yet their impact on the transcriptional profiles of soil microbial communities remains unknown. High alignment of transcript codon frequencies to cellular tRNA anticodon frequencies—also known as codon optimization—has been shown to increase the rate of transcription and translation. Nucleotide composition may also influence transcription due to resource demands for synthesis of specific nucleotides. Here we test the prediction that high codon optimization and low-cost nucleotides will increases the rate of transcription in soil microbial communities. We subjected soils to two separate short-term changes in their environment: an input of labile carbon and a sudden increase in temperature from 20°C to 60°C. From metagenomes we were able to generate a reference set of expected codon frequencies for each taxon. We then used the similarity of transcript codon composition to this reference set to derive codon optimization for transcribed functional genes. We found that inputs of labile carbon resulted in higher level of transcript codon optimization, especially for highly upregulated nitrogen-cycling genes. Nucleotide frequency of differentially regulated transcripts were closely related to codon usage, as opposed to cost of nucleotide synthesis. Surprisingly, we also found a cluster of highly expressed genes with low codon optimization. In both experiments many of these transcripts encoded for proteins related to stress response, such as sporulation and heat shock. We suggest that this is a result of a stress-induced shift in the tRNA pool which better aligns with the codon frequencies of stress-response genes. These results demonstrate the importance of codon optimization for the transcription of functional genes in soil microbial communities and highlight its potential to be used as a metric in predicting the response of soil microorganisms.

INTRODUCTION

Microbial communities in soil are frequently subject to short-term changes in their environment which promote growth and cause disturbance, and determining the factors which dictate the response of soil microbes to these changes is crucial for assessing both their short and long-term function (Shade et al. 2012; Schimel, Balser, and Wallenstein 2007). Although studies have examined how short-term environmental changes influence gene expression in soil microbial communities (Peng, Wegner, and Liesack 2017; Albright et al. 2018; Chuckran, Fofanov, et al. 2021), the influence of nucleotide, codon, and amino acid frequency on the transcription of functional genes remains unknown. These factors are known to relate to gene expression in isolates and represent a potentially underutilized metric in our understanding of how microbial communities may respond to change.

A good alignment of gene codon frequency to the anticodons of the tRNA pool generally increases the rate of both transcription and translation (Plotkin and Kudla 2011) and is commonly referred to as codon optimization. The level of codon optimization of a transcript has important consequences for translation; affecting the rate of elongation, protein folding, initiation, and termination (Yu et al. 2015; M. Zhou et al. 2013; Liu, Yang, and Zhao 2021). Codon optimization is also a strong determinant of mRNA abundnace. A high abundance of optimized codons generally increases mRNA stability (Dressaire et al. 2013; Presnyak et al. 2015) and high codon optimization has also been shown to be independently related to higher levels of transcription (Z. Zhou et al. 2016; Newman et al. 2016). Codon usage has been used to predict growth rates of individual bacterial taxa in microbial communities (Weissman, Hou, and Fuhrman 2021); however, the impact of codon usage bias on the transcription of individual genes has yet to be assessed on a community level.

The selection of specific nucleotides and amino acids can additionally alter gene expression through differential requirement for carbon, energy, and nutrients. Highly expressed proteins tend to require less biosynthetically expensive amino acids (Raiford et al. 2012), and nucleotide selection favors nucleic acids with lower energetic costs of production (W. H. Chen et al. 2016). Nutritional constraints also influence protein amino acid composition for corresponding transporters (Baudouin-Cornu et al. 2001), and may have consequences for their rate of transcription. For example nitrogen limitation has been shown to favor the transcription of transporters with a low requirement of nitrogen (Read et al. 2017). Further, there is a positive relationship between the C:N of codons and the C:N of amino acids (Bragg and Hyder 2004; Shenhav and Zeevi 2020; Sueoka 1961), which would influence nutrient requirement for the production of mRNA. Although the relationship between amino acid content and GC content with nutrient availability has been demonstrated in soil metagenomes (Chuckran, Flagg, et al. 2021), it is not yet known how these factors influence the short-term transcriptional response of soil microbial communities.

In a previous study, we showed that the addition of labile carbon in the form of glucose can rapidly stimulate the transcription of nitrogen cycling genes (Chuckran, Fofanov, et al. 2021). The activity of soil heterotrophic bacteria are generally carbon-limited (Demoling, Figueroa, and Bååth 2007; Hobbie and Hobbie 2013) and the addition of a labile carbon may rapidly stimulate microbial activity. The sudden availability of carbon often causes a shift to nitrogen limitation that causes a rapid immobilization of available nitrogen (Kamble and Bååth 2014). Here we present an analysis of this transcriptional response to a glucose addition with a focus on codon frequency, nucleotide composition, and amino acid composition. We also analyzed metatranscriptomes from a heat-stress experiment, where soils where heated to 60°C

for 30 minutes. We hypothesize that highly and early expressed transcripts will exhibit high codon optimization in response to both the glucose addition and heat stress. We also hypothesize that nucleic and amino acid composition of nutrient transporter transcripts will reflect the nutrient which they are transporting. For example, the GC content and amino acid content of highly upregulated nitrogen transporters will exhibit high carbon and low nitrogen content. Similarly, we expect a high C:N of the predicted amino acid sequences for the entire metatranscriptome when nitrogen becomes limiting. Through this analysis we hope to better understand the predictive power of codon composition, and nucleic and amino acid content in assessing the response potential of soil microbial communities.

METHODS

A more in-depth description of the soil collection and glucose addition experiment can be found in Chuckran et al. 2021.

Soil collection

Soils were collected from the West Virginia Certified Organic Farm (Morgantown, West Virginia, USA, 39.647502°N, 79.93691°W; 243.8 to 475.2 m above sea level) in the fall of 2017. Soils were sampled from plots subject to a four-year conventionally tilled crop cycle of corn, soybean, wheat, and a mix of kale and cowpea, with manure additions every two years and a rye-vetch winter crop cover (Walkup et al. 2020; Pena-Yewtukhiw et al. 2017). Ten cores 0-10 cm in depth were collected from each plot and pooled. Soils were shipped on ice to Northern Arizona University (Flagstaff, Arizona, USA). An equal amount of soil from each plot was pooled. Large roots and debris were manually removed and soil was passed through a 2mm sieve.

Temperature experiment

Pooled soils were distributed in 10 glass Mason jars at 30 g of soil each. Samples were preincubated at ~23°C for 2 weeks. The lid of each jar was briefly opened after preincubation to refresh the headspace of the jar before treatment. Five of the samples were placed in an incubator at 20°C and 5 samples were placed at 60°C for 30 minutes. Soils were then destructively sampled and immediately frozen in liquid N_2 for preservation for nucleic acid extractions.

Glucose addition experiment

A total of 30 g of the pooled soil was distributed among 65 glass Mason jars and allowed to preincubated at ~23°C for 2 weeks. After preincubation, we added 1.6 mL of 0.13 M glucose

(0.7 mg glucose C g^{-1} dry soil) to 60 of the samples. 5 samples which were left untreated as a control. Every 4 hours, the CO_2 concentration in the headspace of each jar was measured and 5 chambers were destructively sampled. During destructive sampling, a portion of the soil was frozen in liquid N for potential nucleotide extractions and another portion was reserved to measure concentrations of NO_3^-, NH_4^+, and microbial biomass.

Metagenomes and metatranscriptomes

RNA and DNA were extracted from 4 samples for each temperature (20°C and 60°C) and for four timepoints from the glucose incubation: 0 (t_0), 8 (t_8), 24 (t_{24}), and 48 h (t_{48}). The RNA and DNA for each sample were extracted using the RNeasy PowerSoil total RNA kit (Qiagen) and RNeasy PowerSoil DNA elution kit (Qiagen), respectively. DNase was removed using an RNase-free DNase set (Qiagen). A Qubit fluorometer (Invitrogen, Carlsbad, CA, USA) was used to assess quantity and NanoDrop ND-1000 spectrophotometer (Nanodrop Technologies, Wilmington, DE, USA) was used to assess sample purity. Samples were shipped to the Joint Genome Institute for sequencing on an Illumina NovaSeq platform (San Diego, CA, USA).

Sequence data was processed by the Joint Genome Institute and a detailed description of the sequencing and bioinformatics pipeline can be found in the associated data release (Chuckran et al. 2020). Raw sequences were QC filtered using BBtools v.38 (Bushnell 2014). Metatranscriptomes were assembled with MEGAHIT v.1.1.2 (Li et al. 2015) and metagenomes were assembled using SPAdes v3.13.0 (Bankevich et al. 2012). Coverage against assembled contigs was determined using BBMap v38 (Bushnell 2014). Contigs were annotated with the IMG Annotation Pipeline v5.0.1 (I.-M. A. Chen et al. 2019; Huntemann et al. 2016).

Calculating of codon optimization and amino acid content

We used three indices to calculate codon optimization: the Codon Adaptation Index (CAI; Sharp and Li 1987), the Frequency of Optimized Codons (FOP; Ikemura 1981), and the Measurement Independent of Length and Composition (MILC; Supek and Vlahoviček 2005). Each of these indices requires a reference set of frequencies from which to calculate codon bias. Often this would be generated using the anticodons of the tRNA pool (Bahiri-Elitzur and Tuller 2021). However, we found that the tRNA annotations from both our metagenomes and metatranscriptomes did not cover a broad range of taxa, severely limiting our estimates of codon optimization. Instead, we used the genomic background codon frequencies of each taxon derived from the metagenomes. Although codon frequencies of the genomic background will not always accurately represent the tRNA pool, they have been successfully used in predicting growth rates in metagenomes (Weissman, Hou, and Fuhrman 2021) and should generally represent the preferred synonymous codon usage in the tRNA pool. BEDTools (Quinlan and Hall 2010) was used to isolate gene sequences in metagenomic contigs from general feature format files generated by the IMG analysis pipeline. Using the taxonomic annotations and read coverage, we determined the relative frequency of each codon for each taxon in our metagenomes. We then removed taxa where the total depth-adjusted codon count was below 750,000, for a total of 720 taxa. We also summarized codon frequencies at the phylum level, recovering codon frequencies for 46 phyla. Taxon-specific codon frequencies were then used to calculate codon indices for each transcript in the metatranscriptomes.

The Codon Adaptation Index (CAI) was calculated by first determining weights for each amino acid from the reference dataset (i.e. the taxa-level codon frequencies from the metagenomes):

[1] $$w_c = \frac{f_c}{max(f_{synonymous\ codons})},$$

where the weight of each codon, w_c, is determined as its frequency in the genome, f_c, divided by the maximum codon frequency of synonymous codons for the encoded amino acid. For example, glutamic acid is encoded by GAA and GAG. If the codon frequencies for a genome were 0.2 for GAA and 0.8 for GAG, the weights would be 0.25 and 1. These reference weights were then applied to each codon in a transcript. Building on the previous example, a transcript with the sequence "GAA GAA GAG" would yield a list of weights [0.25, 0.25, 1] for that transcript. The geometric mean of these weights represents the CAI for that transcript, as shown in equation 2:

[2] $$CAI = (\prod_{c=1}^{L} w_c)^{\frac{1}{L}},$$

where L is the length of the transcript, calculated as the number of codons. Values closer to 1 indicate a high level of codon optimization.

FOP was determined as:

[3] $$FOP = \frac{\#\ of\ optimized\ codons}{Total\ number\ of\ codons},$$

where the optimized codon for each transcript was determined from the reference frequencies for the corresponding taxa. Like CAI, values closer to 1 indicate a higher level of codon optimization. FOP and CAI are generally well correlated (Bahiri-Elitzur and Tuller 2021).

MILC (Supek and Vlahoviček 2005) was calculated by first calculating the bias of each amino acid, M_a, as in equation 4:

[4] $$M_a = 2 \sum O_c \, ln \frac{f_c}{g_c},$$

where O_c is the number of codons, f_c is the observed frequency in the transcript, and g_c is the expected frequency of codon c. MILC is then calculated as in equation 5:

[5] $$MILC = \frac{\sum_a M_a}{L} - C,$$

where L is the length of codons and C is a correction for overall bias in short sequences a calculated in equation 6:

[6] $$C = \frac{\sum_a (r_a - 1)}{L} - 0.5 \,,$$

where r_a is the total number of synonymous codons for amino acid a. MILC values closer to 0 indicate a higher level of similarity to the reference frequencies.

Each index was calculated using a taxon-level reference set of frequencies. Genes with fewer than 80 codons were discarded. Codon frequency and bias calculations were conducted using custom-made Python scripts which are available at https://github.com/PChuckran/Chuckran_dissertation/tree/main/Chapter_4/code

Using the amino acid sequence provided by the IMG annotation pipeline, we summed the number of each amino acid for all genes. Using the chemical formula of the amino acids, we then calculated the total stoichiometric ratios for each gene.

Indices, GC content, and amino acid content were adjusted for read depth, and summarized by either KEGG Orthology (KO) number (Kanehisa and Goto 2000) and taxa; KO number and phylum; or KO number alone. Data was summarized using the python library pandas (McKinney 2011).

Statistical analyses

Differential expression of transcription was determined using DESeq2 (Love, Huber, and Anders 2014) using the total number of gene counts for each gene. A Wald-test was used to determine differences in expression between treatments and a false discover rate (FDR) of < 0.1 indicated significance differences in the expression of a gene between two timepoints or treatments. Differences in codon indices, GC, and amino acid content between treatments were

determined using an analysis of variance (ANOVA). The relationship between log fold change was determined for each treatment using multiple linear regression. All statistical analyses were conducted in R version 4.1.0 (Team 2018) and visualized with the ggplot2 package (Wickham 2016).

RESULTS

Addition of labile carbon caused a rapid immobilization of both carbon and nitrogen, as well as the upregulation of 2549 genes and downregulation of 1273 genes (FDR < 0.1). An increase in temperature from 20°C to 60°C for 30 minutes resulted in the upregulation of 79 genes and the downregulation of 15 genes (FDR < 0.1).

Metatranscriptome-level indices of codon optimization (i.e., the weighted mean of all genes in the metatranscriptome) increased with the addition of glucose across all indices at 8 hours (Tukey's HSD; $p < 0.05$; Fig 1a-c). Codon optimization for differentially regulated genes demonstrated variability between indices with a few notable trends. Upregulated genes had a higher level of codon optimization at t_8 as compared to t_0 (Fig 1d-f). Downregulated genes exhibited similar codon optimization at t_0, t_8, and t_{24}, and had lower codon optimization at t_{48} for CAI and FOP (Fig 1d&e).

The temperature treatment did not influence the level of codon optimization of the whole transcript pool (Fig 1g-i). Although there were some differences in the codon optimization of differentially regulated genes in response to temperature, changes were not consistent between indices (Fig 1j-l).

The average frequency of optimized codons varied over time for individual genes. We highlight changes in codon optimization for genes which are central in nitrogen cycling, particularly nitrogen regulatory genes and the ammonium transport *amt* (Fig. 2). These genes, which were highly upregulated in response to glucose, also tended to have higher transcript codon optimization as compared to t_0. The codon optimization of housekeeping and glutamate dehydrogenase transcripts was less variable and transcripts for glutamate dehydrogenase *gudB* and *rocG* (KO:K00260) demonstrated a decrease in codon optimization over time (Fig. 2).

Although upregulated transcripts were associated with higher level of codon optimization overall (Fig 1), we found a negative relationship between codon optimization and expression (Fig. 3), measured as $\log_2$-fold change (L2FC) from time zero. This relationship was significant at all timepoints ($p < 0.01$) and was most pronounced at t_{24} ($R^2 = 0.17$) and t_{48} ($R^2 = 0.22$). The distribution of upregulated transcripts with respect to codon optimization showed a slight bimodal distribution at t_{24} and t_{48} (Fig. 3). Many of the highly upregulated genes with low codon optimization at t_{48} were related to sporulation or stress response (Fig. 4). In the temperature experiment, highly expressed heat-shock proteins had lower levels of codon optimization at 60°C (Fig. 5).

The GC content of the metatranscriptomes was higher at t_8 than t_0 ($p < 0.05$, Tukey's HSD) with t_{24} and t_{48} overlapping between t_8 and t_0 (Fig. 6a). The GC content of transcripts for the nitrogen transporter *amt* were higher at t_8, t_{24}, and t_{48} than at t_0 ($p < 0.05$; Fig. 6b). The total average amino acid C:N decreased initially 8 h after glucose addition, and subsequently increased for t_{24} and t_{48} (Fig. 6c).

DISCUSSION

Our initial hypotheses were that fast-responding transcripts would reflect both codon optimization and resource stoichiometry. We found evidence for the former; however, we did not find a strong relationship between nucleic or amino acid stoichiometry and transcription for functional genes associated with ammonium transport.

Codon usage

Codon optimization increased after the addition of glucose and was generally higher for upregulated genes. Surprisingly, the level of upregulation (L2FC) and codon optimization were negatively correlated. This was especially pronounced later in the incubation when nitrogen becomes limiting. Although a large majority of upregulated transcripts had high codon optimization, this negative relationship between L2FC and optimization is still counter-intuitive, as we would expect a high level of codon optimization to be associated with higher upregulation. One potential explanation is that this relationship is a result of over-inflated L2FC values for transcripts that are in relatively low abundance. Highly upregulated genes with few transcripts overall tended to have lower levels of codon optimization (Fig. S1). Since the calculated L2FC of transcripts in low abundance is intrinsically more sensitive to small changes, we could be detecting high stochasticity in our calculation of L2FC. However, this explanation still does not explain why transcripts with low codon optimization would be upregulated to begin with.

We suggest that this anomaly is a result of shifting tRNA pools with stress—in this case—nutrient limitation. Stress can alter tRNA pools by increasing the abundance of rare anti-codons and stress-response genes can be especially well optimized to this set of tRNA (Advani and Ivanov 2019). This optimization to the altered tRNA pool allows for more rapid upregulation of

these genes during stress. Since the background codon frequencies for a genome were used to calculate the expected codon composition, our estimates of codon optimization are not specific to the tRNA pool for a given moment in time. A shift in tRNA frequencies towards more rare anticodons could therefore result in the high upregulation of genes with low levels of codon optimization against the genomic background. This mechanism would also explain the higher abundance of these genes at t_{48}, when nitrogen and labile carbon availability has been substantially depleted. Accordingly, when we examine the function of this cluster of transcripts, we find that a considerable portion of these transcripts encode for proteins involved with sporulation (Fig. 4).

This hypothesis is further supported by the results of the heat-shock experiment. Although we did not observe a shift in codon optimization overall, we did find that highly upregulated heat shock proteins, such as *groEL* and *DnaK* (Richter, Haslbeck, and Buchner 2010), showed lower levels of codon optimization after heat-stress (Fig. 5). We believe that this result, in conjunction with the low optimization of sporulation transcripts, provides evidence for the importance of codon optimization in both growth and stress response of bacteria in soil microbial communities.

This work also indicates the potential for codon optimization to be used to predict the response of soil microbial communities to changes in the environment. Codon usage has been previously leveraged to predict growth rates of soil microbes using codon frequencies in ribosomal proteins (Weissman, Hou, and Fuhrman 2021). Our work demonstrates that optimization is additionally important for the transcription of functional genes in soil microbial communities. By calculating codon optimization of genes for specific functions, we may be able to predict how microbial communities respond to change. For example, higher optimization in response to more optimal conditions and lower when stress and limitations increase. This is not

only important for predicting growth rates but could also be used for assessing the capacity of microbial communities to resist disturbance. Resistance and resilience are central to microbial community dynamics (Shade et al. 2012) and the ability to predict community response to disturbance from codon frequencies could be a valuable tool in metagenomic analyses.

GC content and amino acid stoichiometry

We observed an overall increase in GC content which was highly variable between genes and that GC content did not correspond to specific transporters as we hypothesized. For example, if the regulation of transcription of transporters was closely linked to resource stoichiometry, then we would expect a decrease, rather than an increase, in GC content (since the GC base pair is more nitrogen rich than the AT base pair). We believe that the observed change in GC content over time is more likely a reflection of increased codon optimization rather than resource conservation. The mean metagenomic GC content was 64%, whereas the metatranscriptomes ranged from 56-60%. Since the base codon frequencies used to calculate codon optimization were determined from all open reading frames in the metagenome, it stands to reason that an increase in optimization coincides with an increase in GC content.

Amino acid stoichiometry only weakly followed the predicted relationship and never significantly deviated from that of t_0. Although amino acid stoichiometry may be important cost-saving measures for bacteria, they do not appear to be related to the response time of transcription.

CONCLUSIONS

This work revealed several interesting relationships between codon optimization and the expression of functional genes in microbial communities. Community-level codon optimization increased after the addition of glucose, especially for genes encoding for nitrogen transport. This result indicates the importance of codon frequency for the rapid response of soil microbes to changing nutrient availability. Finally, we found that highly upregulated stress response genes had low levels of optimization, which we suggest could be due to shifting tRNA pools with stress. Together, these results demonstrate the importance of codon usage in the response of soil microbes to change and highlight the potential utility of codon usage for predicting the responses of soil microbial communities.

DATA AVAILABILITY

Metagenomes and metatranscriptomes from the glucose addition experiment have been described previously in a data release: https://doi.org/10.1128/MRA.00895-20

Sample data for the metagenomes and metatranscriptomes from the temperature increase experiment, including sequence read archive accession number, IMG identification number, and N_{50} values, can be found in Supplemental Table 1.

Advani, Vivek M., and Pavel Ivanov. 2019. “Translational Control under Stress: Reshaping the Translatome.” *BioEssays* 41 (5): 1900009. https://doi.org/10.1002/BIES.201900009.

Albright, Michaeline B.N., Renee Johansen, Deanna Lopez, La Verne Gallegos-Graves, Blaire Steven, Cheryl R Kuske, and John Dunbar. 2018. “Short-Term Transcriptional Response of Microbial Communities to Nitrogen Fertilization in a Pine Forest Soil.” *Applied and Environmental Microbiology* 84 (15): e00598-18. https://doi.org/10.1128/AEM.00598-18.

Bahiri-Elitzur, Shir, and Tamir Tuller. 2021. “Codon-Based Indices for Modeling Gene Expression and Transcript Evolution.” *Computational and Structural Biotechnology Journal* 19 (January): 2646–63. https://doi.org/10.1016/J.CSBJ.2021.04.042.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. 2012. “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing.” *Journal of Computational Biology* 19 (5): 455–77. https://doi.org/10.1089/cmb.2012.0021.

Baudouin-Cornu, P., Y. Surdin-Kerjan, P. Marlière, and D. Thomas. 2001. “Molecular Evolution of Protein Atomic Composition.” *Science* 293 (5528): 297–300. https://doi.org/10.1126/science.1061052.

Bragg, Jason G., and Charles L. Hyder. 2004. “Nitrogen versus Carbon Use in Prokaryotic Genomes and Proteomes.” *Proceedings of the Royal Society B: Biological Sciences* 271 (SUPPL. 5): 374–77. https://doi.org/10.1098/rsbl.2004.0193.

Bushnell, Brian. 2014. “BBTools Software Package.” 2014. https://sourceforge.net/projects/bbmap/.

Chen, I-Min A, Ken Chu, Krishna Palaniappan, Manoj Pillay, Anna Ratner, Jinghua Huang, Marcel Huntemann, et al. 2019. “IMG/M v.5.0: An Integrated Data Management and

Comparative Analysis System for Microbial Genomes and Microbiomes." *Nucleic Acids Research* 47 (D1): D666–77. https://doi.org/10.1093/nar/gky901.

Chen, Wei Hua, Guanting Lu, Peer Bork, Songnian Hu, and Martin J. Lercher. 2016. "Energy Efficiency Trade-Offs Drive Nucleotide Usage in Transcribed Regions." *Nature Communications* 7 (1): 1–10. https://doi.org/10.1038/ncomms11334.

Chuckran, Peter F., Cody Flagg, Jeffrey Propster, William A. Rutherford, Ella Sieradzki, Steven J. Blazewicz, Bruce Hungate, Jennifer Pett-Ridge, Egbert Schwartz, and Paul Dijkstra. 2021. "Edaphic Controls on Genome Size and GC Content of Bacteria in Soil Microbial Communities." *BioRxiv* 5 (November): 2021.11.17.469016. https://doi.org/10.1101/2021.11.17.469016.

Chuckran, Peter F., Viacheslav Fofanov, Bruce A. Hungate, Ember M. Morrissey, Egbert Schwartz, Jeth Walkup, and Paul Dijkstra. 2021. "Rapid Response of Nitrogen Cycling Gene Transcription to Labile Carbon Amendments in a Soil Microbial Community." *MSystems* 6 (3). https://doi.org/10.1128/MSYSTEMS.00161-21.

Chuckran, Peter F., Marcel Huntemann, Alicia Clum, Brian Foster, Bryce Foster, Simon Roux, Krishnaveni Palaniappan, et al. 2020. "Metagenomes and Metatranscriptomes of a Glucose-Amended Agricultural Soil." *Microbiology Resource Announcements* 9 (44). https://doi.org/10.1128/mra.00895-20.

Demoling, Fredrik, Daniela Figueroa, and Erland Bååth. 2007. "Comparison of Factors Limiting Bacterial Growth in Different Soils." *Soil Biology and Biochemistry* 39 (10): 2485–95. https://doi.org/10.1016/J.SOILBIO.2007.05.002.

Dressaire, Clémentine, Flora Picard, Emma Redon, Pascal Loubière, Isabelle Queinnec, Laurence Girbal, and Muriel Cocaign-Bousquet. 2013. "Role of MRNA Stability during

Bacterial Adaptation." *PLOS ONE* 8 (3): e59059. https://doi.org/10.1371/JOURNAL.PONE.0059059.

Hobbie, John E., and Erik A. Hobbie. 2013. "Microbes in Nature Are Limited by Carbon and Energy: The Starving-Survival Lifestyle in Soil and Consequences for Estimating Microbial Rates." *Frontiers in Microbiology* 4 (November): 324. https://doi.org/10.3389/fmicb.2013.00324.

Huntemann, Marcel, Natalia N. Ivanova, Konstantinos Mavromatis, H. James Tripp, David Paez-Espino, Kristin Tennessen, Krishnaveni Palaniappan, et al. 2016. "The Standard Operating Procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v.4)." *Standards in Genomic Sciences* 11 (1). https://doi.org/10.1186/s40793-016-0138-x.

Ikemura, Toshimichi. 1981. "Correlation between the Abundance of Escherichia Coli Transfer RNAs and the Occurrence of the Respective Codons in Its Protein Genes: A Proposal for a Synonymous Codon Choice That Is Optimal for the E. Coli Translational System." *Journal of Molecular Biology* 151 (3): 389–409. https://doi.org/10.1016/0022-2836(81)90003-6.

Kamble, Pramod N., and Erland Bååth. 2014. "Induced N-Limitation of Bacterial Growth in Soil: Effect of Carbon Loading and N Status in Soil." *Soil Biology and Biochemistry* 74 (July): 11–20. https://doi.org/10.1016/J.SOILBIO.2014.02.015.

Kanehisa, Minoru, and Susumu Goto. 2000. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research* 28 (1): 27–30. https://doi.org/10.1093/NAR/28.1.27.

Li, Dinghua, Chi Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak Wah Lam. 2015. "MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph." *Bioinformatics* 31 (10): 1674–76. https://doi.org/10.1093/bioinformatics/btv033.

Liu, Yi, Qian Yang, and Fangzhou Zhao. 2021. “Synonymous but Not Silent: The Codon Usage Code for Gene Expression and Protein Folding.” *Annual Review of Biochemistry*. https://doi.org/10.1146/annurev-biochem-071320-112701.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12). https://doi.org/10.1186/s13059-014-0550-8.

McKinney, Wes. 2011. “Pandas: A Foundational Python Library for Data Analysis and Statistics.” *Python for High Performance and Scientific Computing*, 1–9. http://pandas.sf.net.

Newman, Zachary R., Janet M. Young, Nicholas T. Ingolia, and Gregory M. Barton. 2016. “Differences in Codon Bias and GC Content Contribute to the Balanced Expression of TLR7 and TLR9.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (10): E1362–71. https://doi.org/10.1073/PNAS.1518976113/-/DCSUPPLEMENTAL.

Pena-Yewtukhiw, Eugenia M., Emily Leslie Romano, Nicole Lynn Waterland, and John H. Grove. 2017. “Soil Health Indicators during Transition from Row Crops to Grass-Legume Sod.” *Soil Science Society of America Journal* 81 (6): 1486–95. https://doi.org/10.2136/sssaj2016.12.0439.

Peng, Jingjing, Carl Eric Wegner, and Werner Liesack. 2017. “Short-Term Exposure of Paddy Soil Microbial Communities to Salt Stress Triggers Different Transcriptional Responses of Key Taxonomic Groups.” *Frontiers in Microbiology* 8 (MAR): 400. https://doi.org/10.3389/FMICB.2017.00400/BIBTEX.

Plotkin, Joshua B., and Grzegorz Kudla. 2011. “Synonymous but Not the Same: The Causes and

Consequences of Codon Bias." *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2899.

Presnyak, Vladimir, Najwa Alhusaini, Ying Hsin Chen, Sophie Martin, Nathan Morris, Nicholas Kline, Sara Olson, et al. 2015. "Codon Optimality Is a Major Determinant of MRNA Stability." *Cell* 160 (6): 1111–24. https://doi.org/10.1016/J.CELL.2015.02.029.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/BIOINFORMATICS/BTQ033.

Raiford, Douglas W., Esley M. Heizer, Robert V. Miller, Travis E. Doom, Michael L. Raymer, and Dan E. Krane. 2012. "Metabolic and Translational Efficiency in Microbial Organisms." *Journal of Molecular Evolution* 74 (3–4): 206–16. https://doi.org/10.1007/S00239-012-9500-9/FIGURES/3.

Read, Robert W, Paul M Berube, Steven J Biller, Iva Neveux, Andres Cubillos-Ruiz, Sallie W Chisholm, and Joseph J Grzymski. 2017. "Nitrogen Cost Minimization Is Promoted by Structural Changes in the Transcriptome of N-Deprived Prochlorococcus Cells." *The ISME Journal* 11: 2267–78. https://doi.org/10.1038/ismej.2017.88.

Richter, Klaus, Martin Haslbeck, and Johannes Buchner. 2010. "The Heat Shock Response: Life on the Verge of Death." *Molecular Cell* 40 (2): 253–66. https://doi.org/10.1016/j.molcel.2010.10.006.

Schimel, Joshua, Teri C. Balser, and Matthew Wallenstein. 2007. "Microbial Stress-Response Physiology and Its Implications for Ecosystem Function." *Ecology* 88 (6): 1386–94. https://doi.org/10.1890/06-0219.

Shade, Ashley, Hannes Peter, Steven D. Allison, Didier L. Baho, Mercè Berga, Helmut Bürgmann, David H. Huber, et al. 2012. "Fundamentals of Microbial Community

Resistance and Resilience." *Frontiers in Microbiology*. Frontiers Research Foundation. https://doi.org/10.3389/fmicb.2012.00417.

Sharp, Paul M., and Wen Hsiung Li. 1987. "The Codon Adaptation Index--a Measure of Directional Synonymous Codon Usage Bias, and Its Potential Applications." *Nucleic Acids Research* 15 (3): 1281. https://doi.org/10.1093/NAR/15.3.1281.

Shenhav, Liat, and David Zeevi. 2020. "Resource Conservation Manifests in the Genetic Code." *Science* 370 (6517): 683–87. https://doi.org/10.1126/science.aaz9642.

Sueoka, Noboru. 1961. "Compositional Correlation between Deoxyribonucleic Acid and Protein." *Cold Spring Harbor Symposia on Quantitative Biology* 26: 35–43.

Supek, Frank, and Kristian Vlahoviček. 2005. "Comparison of Codon Usage Measures and Their Applicability in Prediction of Microbial Gene Expressivity." *BMC Bioinformatics* 6 (1): 1–15. https://doi.org/10.1186/1471-2105-6-182/TABLES/3.

Team, R Core. 2018. "R: A Language and Environment for Statistical Computing." *R Foundation for Statistical Computing. Vienna, Austria.*

Walkup, Jeth, Zachary Freedman, James Kotcon, and Ember M. Morrissey. 2020. "Pasture in Crop Rotations Influences Microbial Biodiversity and Function Reducing the Potential for Nitrogen Loss from Compost." *Agriculture, Ecosystems and Environment* 304 (December). https://doi.org/10.1016/j.agee.2020.107122.

Weissman, Jake L., Shengwei Hou, and Jed A. Fuhrman. 2021. "Estimating Maximal Microbial Growth Rates from Cultures, Metagenomes, and Single Cells via Codon Usage Patterns." *Proceedings of the National Academy of Sciences* 118 (12): e2016810118. https://doi.org/10.1073/pnas.2016810118.

Wickham, Hadley. 2016. "Elegant Graphics for Data Analysis." In *Elegant Graphics for Data*

Analysis, 3–10. https://doi.org/10.1007/978-3-319-24277-4_1.

Yu, Chien Hung, Yunkun Dang, Zhipeng Zhou, Cheng Wu, Fangzhou Zhao, Matthew S. Sachs, and Yi Liu. 2015. "Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-Translational Protein Folding." *Molecular Cell* 59 (5): 744–54. https://doi.org/10.1016/J.MOLCEL.2015.07.018.

Zhou, Mian, Jinhu Guo, Joonseok Cha, Michael Chae, She Chen, Jose M. Barral, Matthew S. Sachs, and Yi Liu. 2013. "Non-Optimal Codon Usage Affects Expression, Structure and Function of Clock Protein FRQ." *Nature 2013 495:7439* 495 (7439): 111–15. https://doi.org/10.1038/nature11833.

Zhou, Zhipeng, Yunkun Danga, Mian Zhou, Lin Li, Chien Hung Yu, Jingjing Fu, She Chen, and Yi Liu. 2016. "Codon Usage Is an Important Determinant of Gene Expression Levels Largely through Its Effects on Transcription." *Proceedings of the National Academy of Sciences of the United States of America* 113 (41): E6117–25. https://doi.org/10.1073/pnas.1606724113.

LIST OF FIGURES

Figure 1:
Changes in codon optimization with glucose addition or temperature increased. Note that values closer to 1 represent high levels of optimization for the Codon Adaptation Index (CAI) and Frequency of Optimized Codons (FOP), and values closer to 0 represent higher levels of optimization for the Measure Independent of Length and Composition (MILC). Codon optimization of all transcripts after the addition of glucose using the FOP (**a)**, CAI (**b**), and MILC (**c**). Changes in codon optimization of KEGG annotated transcripts with color and shape indicating differential expression (FOP, **d**; CAI, **e**; and MILC, **f**). Codon optimization for all transcripts after 30 minutes at 20°C and 60°C (FOP, **g**; CAI, **h**; and MILC, **i)**, and with respects to regulation for KEGG annotated transcripts (FOP, **j**; CAI, **k**; and MILC, **l)**.

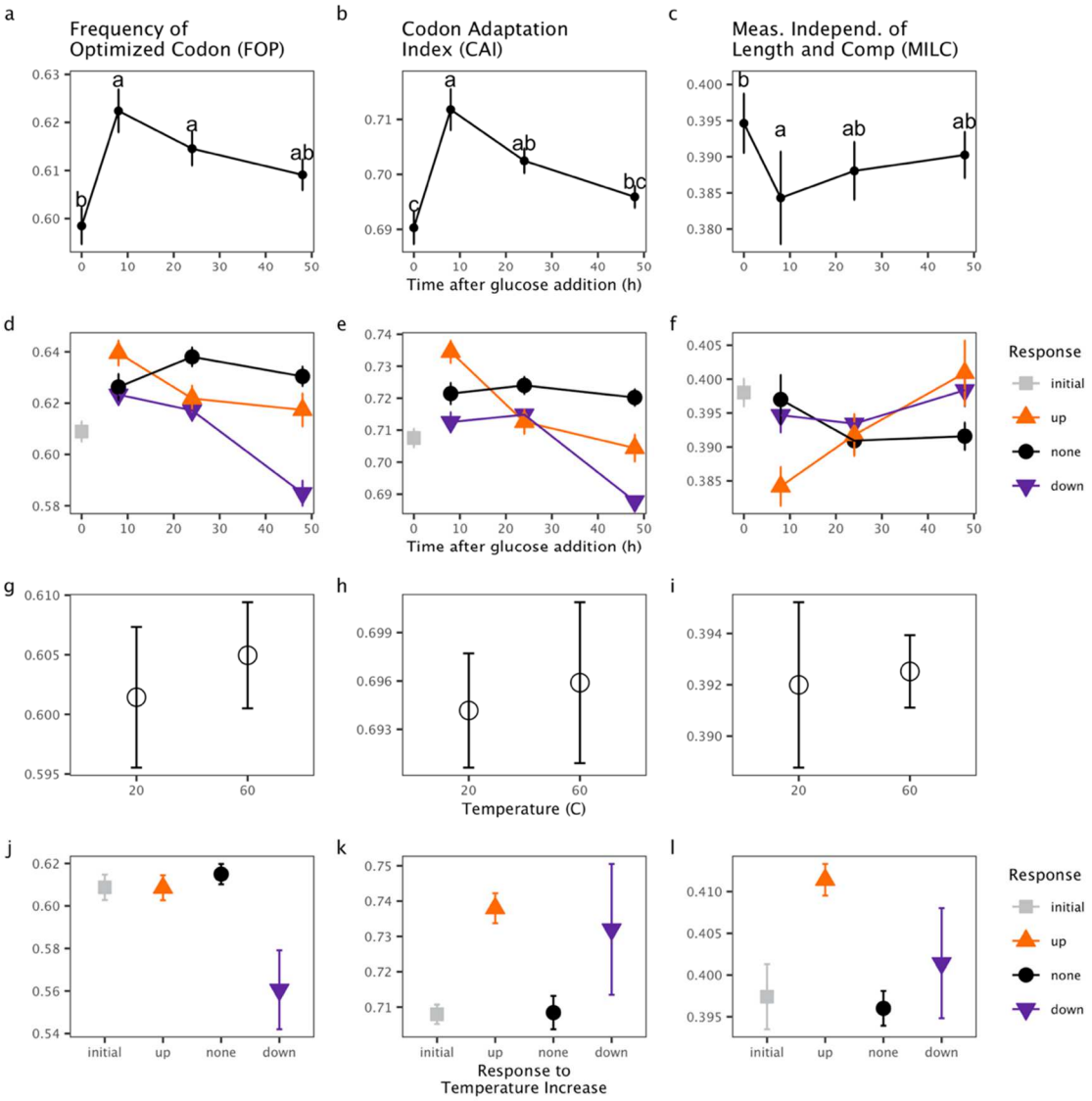


Figure 2

Changes in the frequency of optimized codons (ΔFOP) at 8, 24, and 48 h after glucose addition for ammonium transporter *amt*, glutamate dehydrogenase, glutamine synthase, select housekeeping genes and nitrogen regulatory genes.

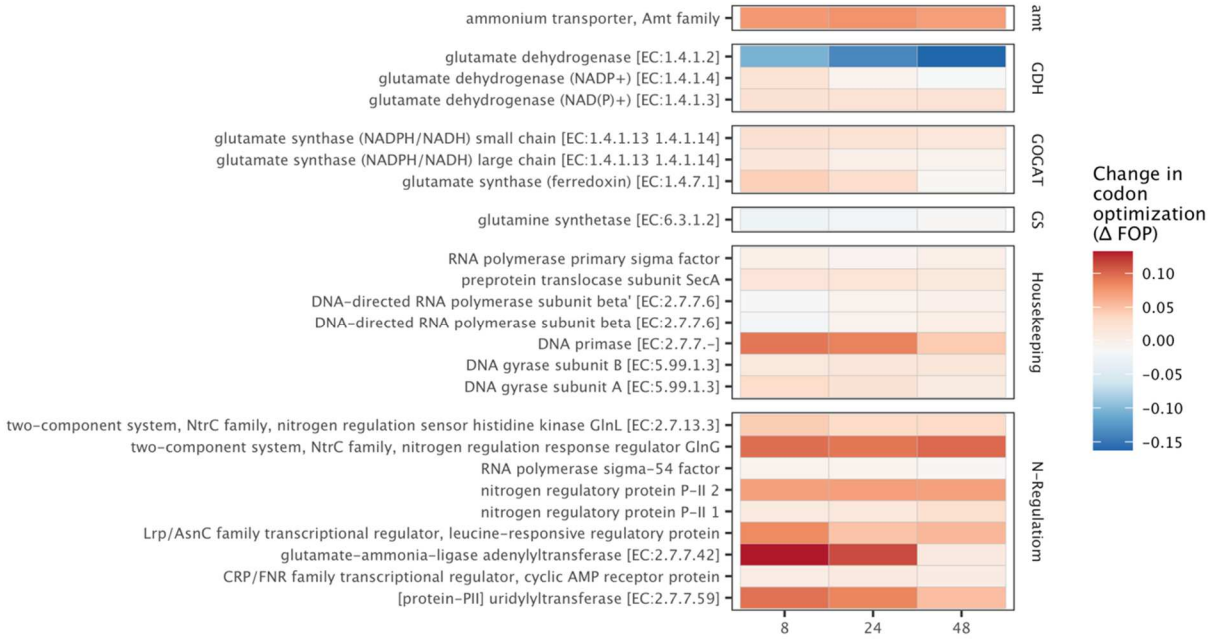


Figure 3
The relationship between the Frequency of Optimized Codons (FOP) and log_2-Fold Change at 8, 24, and 48 hours after glucose addition. Color represents the regulatory response and density distribution curves of FOP are shown above each plot.

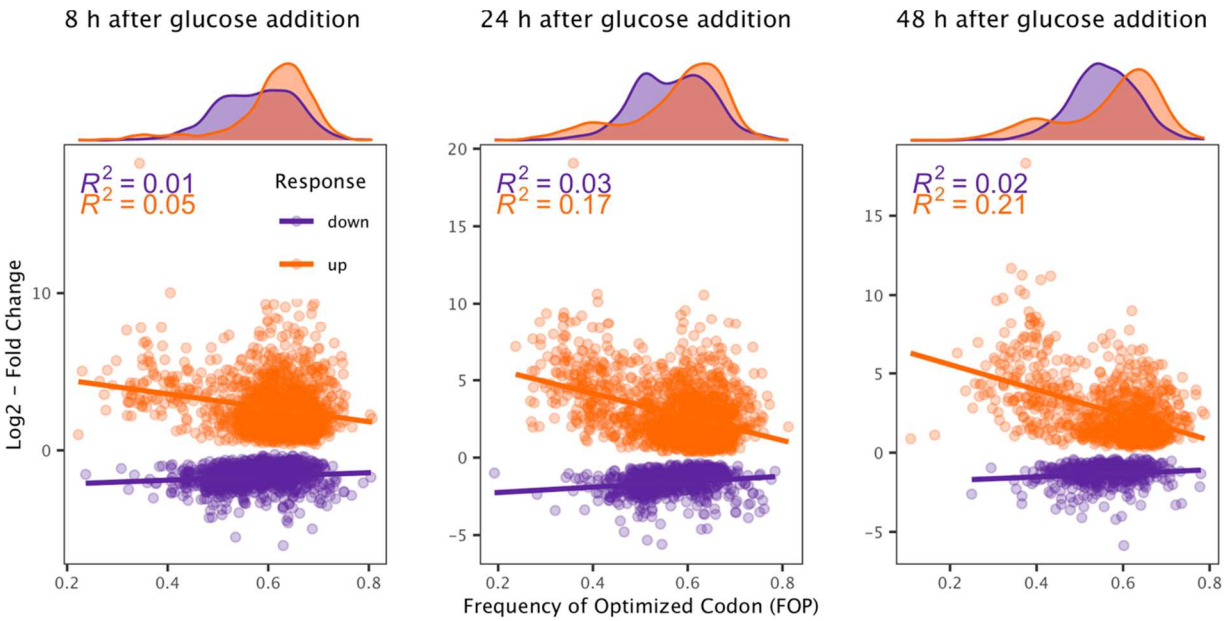


Figure 4
The regulation of KEGG annotated transcripts for sporulation in relation to the Frequency of Optimized Codons (FOP) 48 h after glucose addition. Sporulation genes indicated with red points, with grey points representing all other KEGG annotated genes.

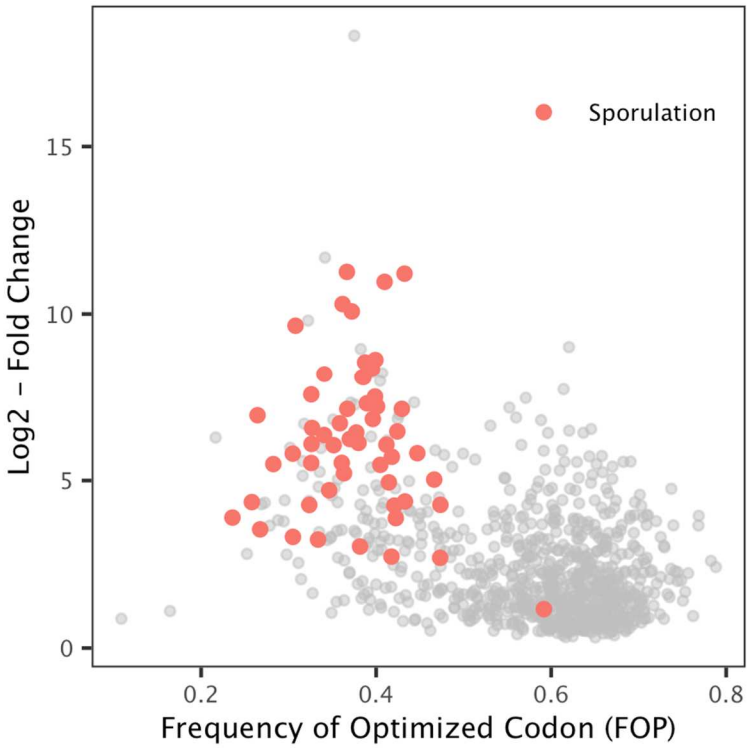


Figure 5

Codon optimization (Frequency of optimized codons; FOP) of upregulated heat-shock proteins.

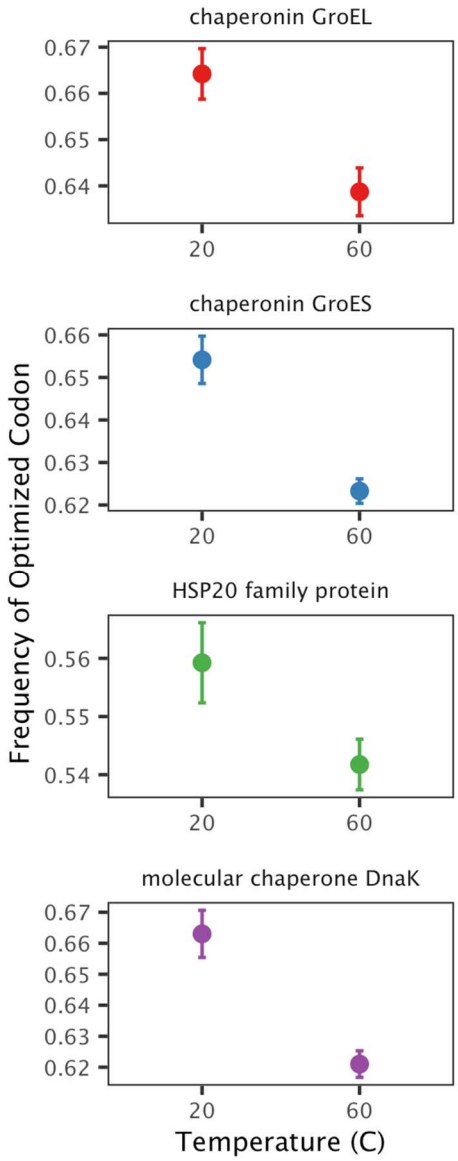


Figure 6
The response of nucleotide and amino acid content in response to the addition of glucose. The GC content after the addition of glucose for all transcripts (**a**) and the ammonium transporter *amt* (**b**). The predicted amino acid carbon to nitrogen ratio (C:N) of all transcripts over time **(c)**.

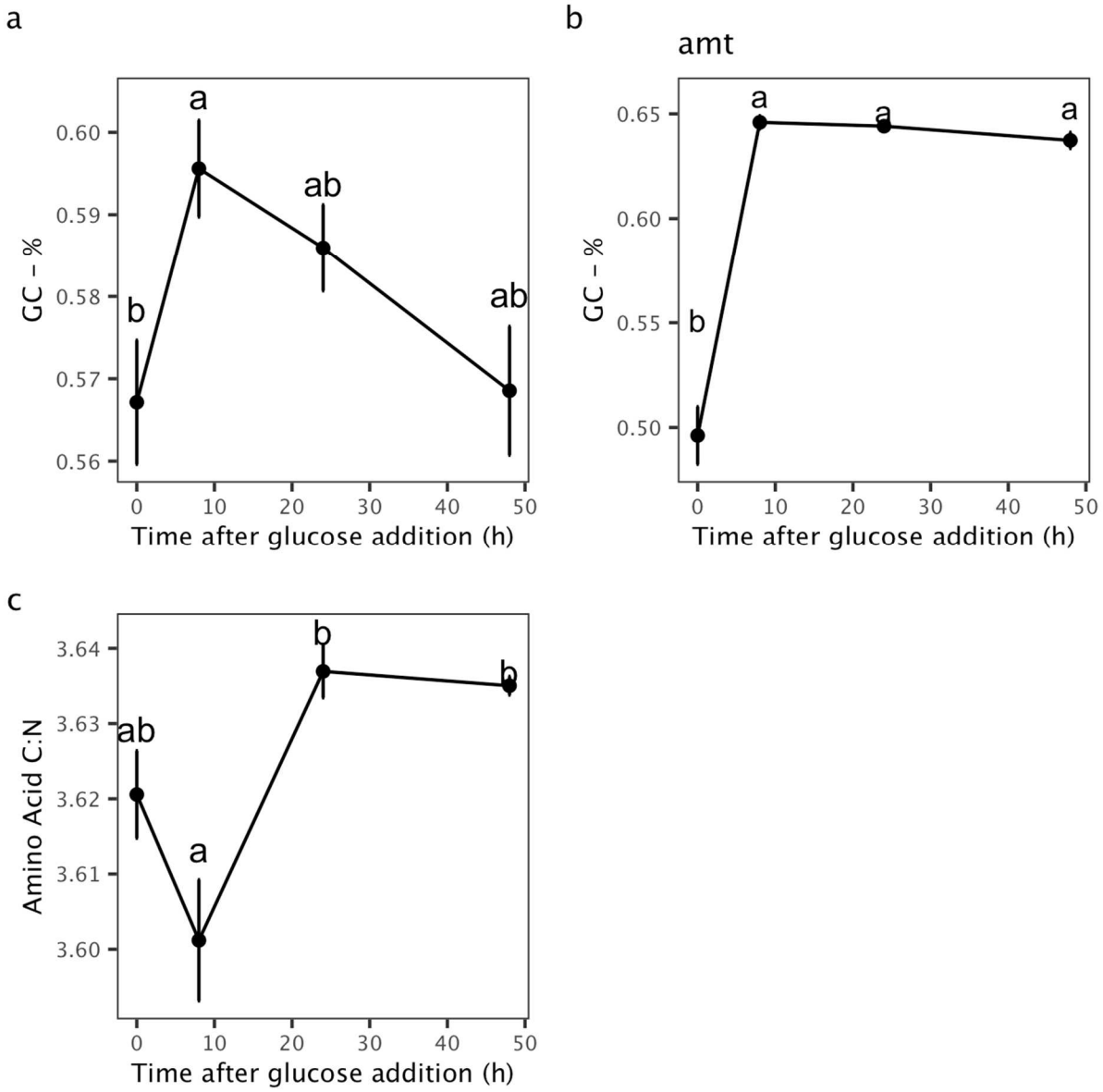
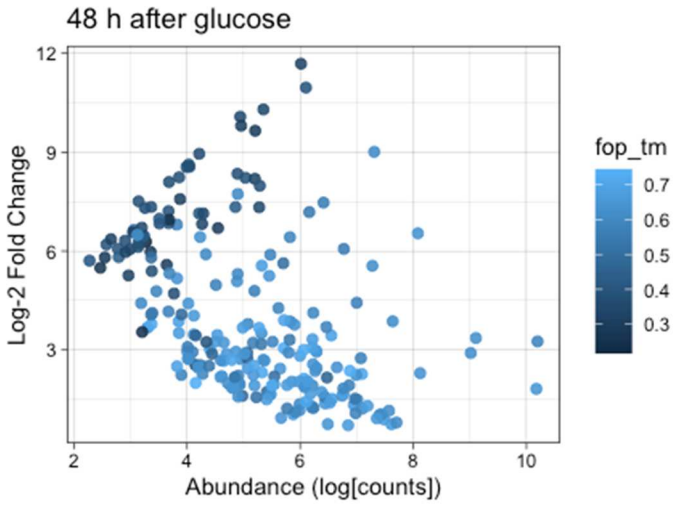


Figure S1
Expression of upregulation genes as it relates to transcript abundance 48 hours after glucose addition. Frequency of Optimized Codons (FOP) indicated by color.



LIST OF TABLES

Supplement Table 1

Sample metadata for metagenomes and metatranscriptomes in temperature incubation experiment.

Temperature	Sample Type	Rep	IMG Genome ID	GOLD Analysis Project ID	NCBI Bioproject Accession	NCBI Biosample Accession	Gene Count	N_{50}
20	Metagenome	1	3300034662	Ga0314783	PRJNA570395	SAMN12813568	762561	1288729
20	Metagenome	2	3300034663	Ga0314784	PRJNA570396	SAMN12814693	747033	880870
20	Metagenome	3	3300034479	Ga0314785	PRJNA570397	SAMN12813706	108601	1913854
20	Metagenome	4	3300034661	Ga0314782	PRJNA570394	SAMN12813049	896719	505047
60	Metagenome	1	3300034664	Ga0314786	PRJNA570398	SAMN12812626	792202	1170933
60	Metagenome	2	3300031547	Ga0310887	PRJNA518696	SAMN10864355	5152793	1009918
60	Metagenome	3	3300031943	Ga0310885	PRJNA539707	SAMN11532793	4298207	1282903
60	Metagenome	4	3300032179	Ga0310889	PRJNA539708	SAMN11532414	3727730	1194338
20	Metatranscriptome	1	3300031913	Ga0310891	PRJNA539710	SAMN11532358	2019847	232158
20	Metatranscriptome	2	3300031538	Ga0310888	PRJNA518697	SAMN10864145	5457649	38430
20	Metatranscriptome	3	3300034659	Ga0314780	PRJNA570392	SAMN12814267	930641	243448
20	Metatranscriptome	4	3300031562	Ga0310886	PRJNA518695	SAMN10864146	5429091	145253
60	Metatranscriptome	1	3300034660	Ga0314781	PRJNA570393	SAMN12814181	583797	282854
60	Metatranscriptome	2	3300031944	Ga0310884	PRJNA539706	SAMN11532342	5024730	174757
60	Metatranscriptome	3	3300034665	Ga0314787	PRJNA570399	SAMN12813567	480705	271809
60	Metatranscriptome	4	3300032075	Ga0310890	PRJNA539709	SAMN11532103	8416504	230787

DISCUSSION OF RESULTS AND CONCLUSIONS

In the introduction, I discuss how growth, stress, and disturbance are the backbone of many of our current frameworks for microbial ecology (Lauro et al. 2009; Malik et al. 2020; Grime 1977; Fierer 2017) and the chapters of this dissertation focus on specific attributes associated with one or more of these themes. To close this dissertation, I will put our findings in the context of Grime's 1977 CSR framework and the YAS framework posed by Malik et al. 2020. Grime's CSR framework (Figure 1) includes: (C) competitors which can effectively compete for resources; (S) stress tolerators which can withstand environmental stress and disturbance; and (R) ruderal strategists which can recover rapidly to disturbance. Similarly, the YAS framework proposes three similar dimensions (Figure 1): (Y) high yield strategists, which invests heavily in rapid response and central metabolism; (A) nutrient acquisition strategists, with high investment in extracellular enzymes and competition for resources; and (S) stress tolerators. We present our results against both frameworks in Figure 1 as described below.

In Chapter 2 we find that certain taxa respond to inputs of labile carbon through rapid transcription of nitrogen cycling genes. In the YAS framework we could describe this response as a high yield strategy, whereas in the CSR framework this might be considered a competitor. Since the sudden input of a large amount of a limiting nutrient could be considered a disturbance from stasis, ruderal strategists may also have an advantage. Chapter 5 showed that codon usage is related to speed and direction of transcriptional responses for a given gene, and we could accordingly add codon optimization to these life-strategies. The results from that analysis also indicated that rapid stress-response may also be driven by codon optimization to a stress-induced

tRNA pool. The flexibility of the tRNA frequency and corresponding codon alignment in stress-response genes may therefore be part of stress-tolerator strategists.

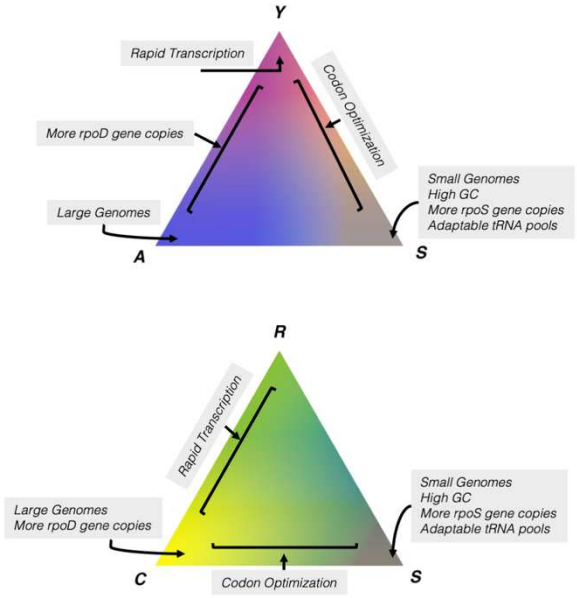
Chapter 3 and 4 examined the biogeographic distribution of genomic traits. We found that soil microbial communities in nutrient limited conditions tended to have smaller genomes with a higher GC content, lower amino acid C:N, and a greater abundance of stress response sigma factor gene *rpoS*. We could consider these traits to be part of a stress-tolerator life strategy. In contrast, communities in low pH but high carbon environments had larger genomes with a lower GC content, and a high abundance of genes for the housekeeping sigma factor *rpoD*. These traits might therefore be part of a nutrient acquisition or competitor strategy.

We do find shortcomings in our ability to classify our results in the context of these frameworks and other frameworks like them. For example, extremely low pH exerts unique stress in comparison to other forms of stress, such as drought. Both pressures would be considered a form of stress, yet they select for very different traits. Ultimately this is the inherit limitation that coincides with attempting to fit any large-scale problem along three axes. Although we can certainly group microbial life into a greater number of dimensions, that comes at the cost of interpretability or, in the extreme, overfitting. These are the shortcomings of using frameworks for describing highly complex systems and it is important to acknowledge that these frameworks can be simultaneously highly useful and deeply imperfect.

Still, the results described in this dissertation point towards fundamental genomic attributes associated with the life-strategies of soil microbes. We show that transcription rate, codon usage, and genomic traits all play important roles in dictating fundamental responses of soil microbes, and we demonstrate how these processes can be detected in short temporal windows and across continental scales. These attributes and their associated mechanisms shed

light onto how omics can be used in assessing soil microbial communities and further our understanding of belowground life.

FIGURE 1:
The results from this dissertation in the context of the YAS framework (top), and the CSR framework (bottom).



REFERENCES

Fierer, Noah. 2017. “Embracing the Unknown: Disentangling the Complexities of the Soil Microbiome.” *Nature Reviews Microbiology* 15 (10): 579–90. https://doi.org/10.1038/nrmicro.2017.87.

Grime, J P. 1977. “Evidence for the Existence of Three Primary Strategies in Plants and Its Relevance to Ecological and Evolutionary Theory.” *The American Naturalist* 111 (982): 1169–94. https://doi.org/10.1086/283244.

Lauro, Federico M, Diane McDougald, Torsten Thomas, Timothy J Williams, Suhelen Egan, Scott Rice, Matthew Z DeMaere, et al. 2009. “The Genomic Basis of Trophic Strategy in Marine Bacteria.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (37): 15527–33. https://doi.org/10.1073/pnas.0903507106.

Malik, Ashish A., Jennifer B.H. Martiny, Eoin L. Brodie, Adam C. Martiny, Kathleen K. Treseder, and Steven D. Allison. 2020. “Defining Trait-Based Microbial Strategies with Consequences for Soil Carbon Cycling under Climate Change.” *ISME Journal* 14 (1): 1–9. https://doi.org/10.1038/s41396-019-0510-0.