ECOLOGICAL STRATEGIES OF SOIL BACTERIA AND ARCHAEA

by

TESS BREWER

B.S., Florida State University, 2012

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirement for the degree of
Doctor of Philosophy
Department of Molecular, Cellular, and Developmental Biology
2019

This thesis entitled:
Ecological Strategies of Soil Bacteria and Archaea
written by Tess Elizabeth Brewer
has been approved for the
Department of Molecular, Cellular, and Developmental Biology

Dr. Noah Fierer

Dr. Kenneth Krauter

Date______________

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Brewer, Tess Elizabeth (Ph.D., Molecular, Cellular, and Developmental Biology)
Ecological Strategies of Soil Bacteria and Archaea
Thesis directed by Professor Noah Fierer

ABSTRACT

Soil is essential for much of life on earth. Microbes are ubiquitous in this environment – billions of microbial cells can occupy one gram of soil. Soil microbes participate in carbon sequestration, nutrient cycling, and soil formation - all critical ecosystem processes, yet are poorly understood. A key factor in this knowledge gap is the low proportion of cultivated soil microbes – by one estimate 1/3 of soil dwelling bacteria and archaea do not have a cultured representative of their phylum. In my thesis research, I have studied the bacteria and archaea that live in soil using culture-independent techniques; specifically studying the unique ecological strategies they employ to excel in what can be a challenging habitat. First, I described how microbial communities change with soil depth; I found that as soil depth increases soil microbes become even more mysterious - candidate phyla and uncultured groups flourish in the low nutrient environment of deep soils. Here, I assembled two genomes from the candidate phylum AD3 and describe the strategies it employs to survive in deep soils. Second, I examined one particular soil bacterium - *Ca.* Udaeobacter copiosus. I show that *Ca.* U. copiosus is incredibly abundant and widespread in soils across the globe, all while relying on a reduced genome with many putative auxotrophies. This observation stands in contrast to prevailing theories that to succeed in soil, bacteria and archaea must possess vast metabolic versatility to take advantage of the diverse, yet limited nutrient sources characteristic of soil. Lastly, I describe a rearrangement of the rRNA operon where the 16S and 23S rRNA genes are "unlinked" and transcribed separately. I show that this rearrangement is common in many environmental bacteria and archaeal groups, and is especially widespread in soil - in one sample 41% of rRNA genes were unlinked. Together, these studies shed a measure of light on the uncultivated majority dwelling in soil, showing that uncultured environmental taxa adopt unique strategies to succeed in this environment, and in some cases harbor biology that stands apart from what we have learned from model organisms like *Bacillus subtilis* and *Escherichia coli*.

ACKNOWLEDGEMENTS

There are a lot of people I would like to acknowledge: I would first like to thank my three primary academic mentors, Kay Jones, Paul Carini, and of course Noah Fierer. They all supported me in different ways and made me feel that I had something of value to contribute to the subjects I work on. They also wrote approximately a bazillion recommendation letters for me over the years. I want to thank all the members of the Fierer lab as well—you know it's a great group when the only major dramatic issue within 5 years was when someone ate someone else's burrito by mistake. I would like to also thank my family—my parents worked hard and moved across the country to give my siblings and I more opportunities, and I think that has worked out okay. Thanks to my sisters, Gwynn and Kara, and my friends Stephen and Erin, for never failing to send me positive psychic energy for important talks and interviews (in jest, we're not psychos okay). Thanks to my friend and peer Wenke, who came to Boulder on a research stay and wormed her way into my life by not-so-subtly hinting that she would like to be invited to a board game night. My life would be much less interesting if I'd never met you (but maybe you should learn to read social cues). Thanks to my cat for tolerating my fawning adoration and my dog for accompanying me on many hikes, even if his pace has slowed with the arrival of his golden years. Lastly, thanks to Mike, who has always supported me and whose opinion I value more than anyone's. Unless the subject is raisins, or nachos. You're an amazing person and you can't argue because this is an acknowledgement, not a conversation. But I really think you should give nachos another chance.

TABLE OF CONTENTS

FIGURES

CHAPTER I
INTRODUCTION AND OVERVIEW

Introduction

Until recently, microbiology research was largely restricted to the fraction of microbial cells capable of being cultured in laboratories. The proportion of cells in an environment that can be cultured is often quite variable - in some environments, like the human body, most of the population can be recapitulated in culture (45-97% of the bacteria and archaea in the human body belong to a genera with a cultured representative (Lloyd et al., 2018). However, the proportion of cultivable cells in soil is a more sobering figure - only 18% of bacteria and archaea in soil belong to genera with a cultured representative (Lloyd et al., 2018). In fact, on average, 1/3 of the bacteria and archaea in soil do not even have a cultured representative of their phylum (as of 2018; Lloyd et al., 2018).

As a result, our understanding of the bacteria and archaea that live in soil is often restricted to fast-growing, copiotrophic taxa. Copiotrophic taxa can easily make use of labile forms of carbon such as glycine and sucrose (Goldfarb et al., 2011), and as a result are often straightforward to culture. Recently, advances in sequencing and bioinformatics have led to reconstruction of genomes independent of cultured isolates, allowing us a peek into the genomes of the uncultured majority (Hug et al., 2016; Rinke et al., 2013). Making use of these genomes, a new tree of life recently demonstrated how far culture collections lag behind true microbial diversity - 68 bacterial and archaeal phyla were without a cultured representative (Hug et al., 2016). These uncultured microbes not only harbor information key to understanding geochemical processes in soil (Yuan et al., 2012; Hayatsu et al., 2008), but also practical information with direct impacts on human health. For example, recent work has shown that a wide variety of genes for the biosynthesis of novel secondary metabolites lurk in the genomes of uncultured Acidobacteria, Verrucomicrobia, Gemmatimonadetes, and Rokubacteria (Crits-Christoph et al., 2018). Indeed, one of the newest antibiotics - teixobactin - was isolated from an uncultured soil bacterium (Ling et al., 2015).

My dissertation research describes these mysterious bacteria and archaea that make up the majority of microbes in soil using culture-independent methods. First, I have shown that soil microbial communities vary consistently with depth across large geographic distances - as soil depth increases nutrients plummet and novel uncultured soil bacteria and archaea generally become more common. I assembled two genomes from an uncultured candidate phylum (Dormibacteraeota, previously AD3) that becomes more abundant with depth and showed that members of this group appear to use ecological strategies tailored to their nutrient poor environment, including spore formation and scavenging of trace CO^2 for sustenance. Secondly, I show that a large genome is not a prerequisite to success in the soil environment. Generally, taxa with larger genomes have more metabolic versatility, a feature thought to enhance survival and niche space in the soil environment, where resources can be diverse, but sparse (Konstantinidis and Tiedje, 2004). Using the uncultured Verrucomicrobia *Candidatus* 'Udaeobacter copiosus' as an example, I show that bacteria with small genomes can be just as ubiquitous in soil, but may be more difficult to culture - likely due to outsourcing of metabolite synthesis to the environment (Giovannoni et al., 2014). Lastly, I discuss an unusual arrangement of rRNA genes where the 16S and 23S are "unlinked" and no longer located in the same operon. I show that unlinked rRNA operons are more widespread in bacteria and archaea than previously assumed, especially among uncultured environmental populations. This unusual arrangement appears to be quite common in soils (41% of rRNA genes in one soil) and suggests that the biology of uncultivated populations does not necessarily conform to the traditional paradigms derived from model organisms like *Escherichia coli* and *Bacillus subtilis.*

Thesis Overview

Chapter II: Uncultured oligotrophic microbes dominate subsurface soils. Resource availability decreases dramatically as you descend through a soil profile – the concentrations of available carbon and nitrogen plummet with depth, as do microbial biomass levels. The bacteria and archaea that live in subsurface soils face a much more challenging resource landscape than those living at the surface. In this chapter, I describe aspects of microbial communities that appear to change consistently with depth regardless of geographic location. I found that the proportion of poorly studied bacteria and archaea rises with depth - candidate phyla and uncultivated groups abound in deep soils. To explore these understudied groups in more detail, I assembled two genomes from members of candidate phylum AD3 from metagenomic sequences and discuss the specific strategies members of this phylum may employ to attain high abundance in deep soils.

Chapter II is adapted from Brewer et al 2019 (fingers crossed). This project was a huge collaborative effort, with samples originating from all 10 Critical Zone Observatories (CZOs) across the US. While there are twenty-five co-authors for this paper, their contributions primarily involved logistics, sample collection and processing, DNA extraction, and soil characterization. I led the amplicon and shotgun metagenomic analyses, along with all bioinformatics and data analyses. I assembled and binned all genomes and performed the PMA spore-selection lab work. Noah Fierer and I wrote the manuscript, with input from all co-authors.

Chapter III: Genome reduction in an abundant and ubiquitous soil bacterium, 'Candidatus Udaeobacter copiosus'. In previous soil studies from the Fierer lab (Ramirez et al., 2014; Fierer et al., 2013; Leff et al., 2015; Fierer et al., 2012; Bergmann et al., 2011), one group of bacteria was almost always found to be abundant in soil ecosystems, regardless of geographic location or soil type - Verrucomicrobia group DA101. This group, while one of the most abundant groups of bacteria in soil, had previously only been described through environmental amplicon sequencing. In order to understand the key to its success in soils, we examined >1000 soils encompassing many soil types from across the globe and found that DA101 was within the top ten most abundant groups of bacteria in over 70% of the soils we analyzed. We leveraged this abundance to assemble a near complete genome from metagenomic data - a difficult feat in soil, where bacterial and archaeal hyper-diversity often confounds assemblers (Howe et al., 2014). We named the organism whose genome we assembled *Candidatus* Udaeobacter copiosus and discovered it has a significantly smaller genome than most soil organisms (~2.8 Mbp versus the average 4.74Mbp; Raes et al., 2007). Because of the high degree of heterogeneity in soils, both in terms of environmental conditions and the quality and quantity of available nutrients, it was previously assumed that soil microorganisms must be capable of metabolizing a broad array of substrates and quickly adapting to changing environmental conditions to succeed in soil, resulting in large average genome sizes (Konstantinidis and Tiedje, 2004; Barberán et al., 2014). *Ca.* U. copiosus is interesting because it does not possess the diverse metabolic capability of other soil organisms - in fact it appears to be auxotrophic for many amino acids. By reducing the components it must directly synthesize, *Ca.* U. copiosus is able to eliminate corresponding biosynthetic machinery, reducing the size of its genome and the baseline level of energy needed to sustain its cells.

This chapter is adapted from Brewer et al 2016. I personally performed all analyses tracking *Ca.* U. copiosus abundance through >1000 soils samples, reconstructed near-full length Verrucomicrobia 16S rRNA sequences with EMIRGE, constructed the Verrucomicrobia phylogeny, and analyzed the metabolic pathways of the *Ca.* U. copiosus genome. The *Ca.* U. copiosus genome was assembled and binned by Kim M. Handley. Noah Fierer, Paul Carini, and I wrote the paper.

Chapter IV: Unlinked rRNA genes are widespread among environmental bacteria and archaea. When a molecule is crucial to the function of a cell, its components are usually highly resistant to change and evolutionarily constrained. For example, textbooks teach that the RNA components of the ribosome (rRNA) are organized into a single operon conserved across all prokaryotic life. In reality, there are some prokaryotes that do not share this canonical rRNA order; their 16S and 23S rRNA genes are separated and referred to as "unlinked". These prokaryotes are under selection to minimize detrimental mutations to their genomes and even possess genetic machinery capable of correcting rogue genome rearrangements. Unlinked rRNA genes are relatively common in soil - roughly 40% of prokaryotic rRNA in some soils is unlinked, a much higher fraction than in other environments (0% in the human gut). Additionally, certain ubiquitous and abundant soil bacteria (*Candidatus* Udaeobacter copiosus of the phylum Verrucomicrobia) also possess unlinked rRNA genes. These two facts imply that unlinked rRNA genes may be advantageous in soil.

In this chapter, I explore which taxa have unlinked rRNA genes using genome databases and long-read sequencing technology, along with the phylogenetic distribution and genomic attributes associated with this trait. I also discuss potential evolutionarily advantages of this arrangement. I performed all analyses associated with this project, leveraging existing sequence data from public databases.

CHAPTER II

UNCULTURED OLIGOTROPHIC MICROBES DOMINATE SUBSURFACE SOILS

Adapted from:

Brewer TE, et al. (2019) Ecological and genomic attributes of novel bacterial taxa that dominate subsurface soil horizons. (Submitted)

Abstract

While most bacterial and archaeal taxa living in surface soil horizons remain undescribed, this problem is exacerbated in deeper soils owing to the unique oligotrophic conditions found in the subsurface. Additionally, previous studies of soil microbiomes have focused almost exclusively on surface soils, even though the microbes living in deeper soils also play critical roles in a wide range of biogeochemical processes. We examined soils collected from 20 distinct profiles across the U.S. to characterize the bacterial and archaeal communities that live in subsurface soils and to determine whether there are consistent changes in soil microbial communities with depth across a wide range of soil and environmental conditions. We found that, irrespective of location, bacterial and archaeal diversity decreased with depth, as did similarity of microbial communities to those found in surface horizons. We observed five phyla that consistently increased in relative abundance with depth across our soil profiles: Chloroflexi, Nitrospirae, Euryarchaeota, and candidate phyla GAL15 and AD3. Leveraging the unusually high abundance of AD3 at depth, we assembled genomes representative of this candidate phylum and identified traits that are likely to be beneficial in low nutrient environments, including the synthesis and storage of carbohydrates, the potential to use carbon monoxide (CO) as a supplemental energy source, and the ability to form spores. Together these attributes likely allow members of the candidate phylum AD3 to flourish in deeper soils and provide insight into the survival and growth strategies employed by the microbial taxa that thrive in oligotrophic soil environments.

Introduction

Subsurface soils often differ from surface horizons with respect to their pH, texture, moisture levels, nutrient concentrations, clay mineralogy, pore networks, redox state, and bulk densities. Globally, the top 20 cm of soil contains nearly five times more organic carbon (C) than soil in the bottom 20 cm of meter-deep profiles (Jobbágy and Jackson, 2000). In addition, residence times of organic C pools are typically far longer in deeper soil horizons (Balesdent et al., 2018), suggesting that much of the soil organic matter found in the subsurface is not readily utilized by microbes. Unsurprisingly, the strong resource gradient observed through most soil profiles is generally associated with large declines in microbial biomass (Schütz et al., 2010; Eilers et al., 2012; Fierer et al., 2003; Blume et al., 2002; Spohn et al., 2016; Stone et al., 2014); per gram soil, microbial biomass is typically one to two orders of magnitude lower in the subsurface than surface horizons (Eilers et al., 2012; Blume et al., 2002; Spohn et al., 2016). Although microbial abundances in deeper soils are relatively low on a per gram soil basis, the cumulative biomass of microbes inhabiting deeper soil horizons can be on par with that living in surface soils, owing to the large mass and volume of subsurface horizons (Fierer et al., 2003; Schütz et al., 2010). Moreover, those microbes living in deeper horizons can play important roles in mediating a myriad of biogeochemical processes, including processes associated with soil C and nitrogen (N) dynamics (Kramer and Gleixner, 2008; Banning et al., 2015), soil formation (Oh and Richter, 2005), iron redox reactions (Fimmen et al., 2008; Hall et al., 2016), and pollutant degradation (Schwarz et al., 2018).

Given that soil properties typically change dramatically with depth, it is not surprising that the composition of soil microbial communities also generally changes with depth through a given profile (Eilers et al., 2012; Fierer et al., 2003; Will et al., 2010; Blume et al., 2002; Kramer et al., 2013; Stone et al., 2014). In some cases, the differences observed in microbial communities with depth through a single soil profile can be large enough to be evident even at the phylum level of resolution. For example, both Chloroflexi (Will et al., 2010; Tas et al., 2014) and Nitrospirae (Will et al., 2010) may increase in relative abundance with depth. However, while previous work suggests that particular taxa can be relatively more abundant in deeper soils, it is unclear if such patterns are consistent across distinct soil and ecosystem types. We hypothesized that there are specific groups of soil bacteria and archaea that are typically rare in surface horizons, but more abundant in deeper soils. Taxa that are proportionally more abundant in deeper soil horizons likely have slow-growing, oligotrophic life history strategies due to the lack of disturbance at depth and the low resource conditions typical of most deeper soil horizons (Fierer, 2017). Likewise, we expect deeper soils to harbor higher proportions of novel and undescribed microbial lineages given that oligotrophic taxa are typically less amenable to in vitro, cultivation-based investigations (Vartoukian et al., 2010).

We designed a comprehensive study to investigate how soil bacterial and archaeal communities change with soil profile depth, to identify taxa that are consistently more abundant in deeper horizons, and to determine what life history strategies enable these taxa to thrive in the resource-limited conditions typical of most subsurface horizons. We collected soil samples at 10-cm increments from 20 soil profiles representing a wide range of ecosystem types throughout the U.S., with most of the profiles sampled to one meter in depth. We examined the bacterial and archaeal communities of these soil profiles by pairing amplicon 16S rRNA gene sequencing with shotgun metagenomic sequencing on a subset of samples. We found that deeper soil horizons typically harbored more undescribed bacterial and archaeal lineages, and we identified specific phyla (including AD3, GAL15, Chloroflexi, Euryarchaeota, and Nitrospirae) that consistently increased in relative abundance with depth across multiple profiles. Moreover, we found one candidate phylum (AD3) to be particularly abundant in deeper soil horizons with low organic C concentrations. From our metagenomic data, we were able to assemble genomes from representative members of this candidate phylum and document the life history strategies, including low maximum growth rates and spore-forming potential, that are likely advantageous under low resource conditions.

Results and Discussion

Sample descriptions and soil properties linked to soil depth

We collected soils from a network of 10 current and former Critical Zone Observatories (CZOs) located across the U.S. (Figure 2.1a) that span a broad range of hydrogeological provinces, soil orders, and ecosystem types, including tropical forest, temperate forest, grassland, and cropland sites. Soils were sampled from two distinct profiles per CZO for a total of 20 different soil profiles. Soils were collected from the first meter (where possible) of freshly excavated profiles, sampling at 10 cm increments and focusing on mineral soil horizons only (O horizons, if present, were not sampled). Together, this collection effort yielded 179 individual soil samples collected across sites with a wide range of different climatic conditions (e.g., mean annual temperatures ranging between 5 - 23 °C and mean annual precipitation ranging from 26 - 402 cm y^{-1}). The sampled profiles ranged from poorly developed Entisols and Inceptisols to highly developed Oxisols and Ultisols (as per the U.S. Soil Taxonomy system), with the samples reflecting an extremely broad range of soil properties. For example, in the 0-10 cm depth increment, soil pH ranged from 3.3 to 9.8, organic carbon concentrations spanned 1.3% to 21.6%, and texture from 0% to 45% silt + clay across the profiles.

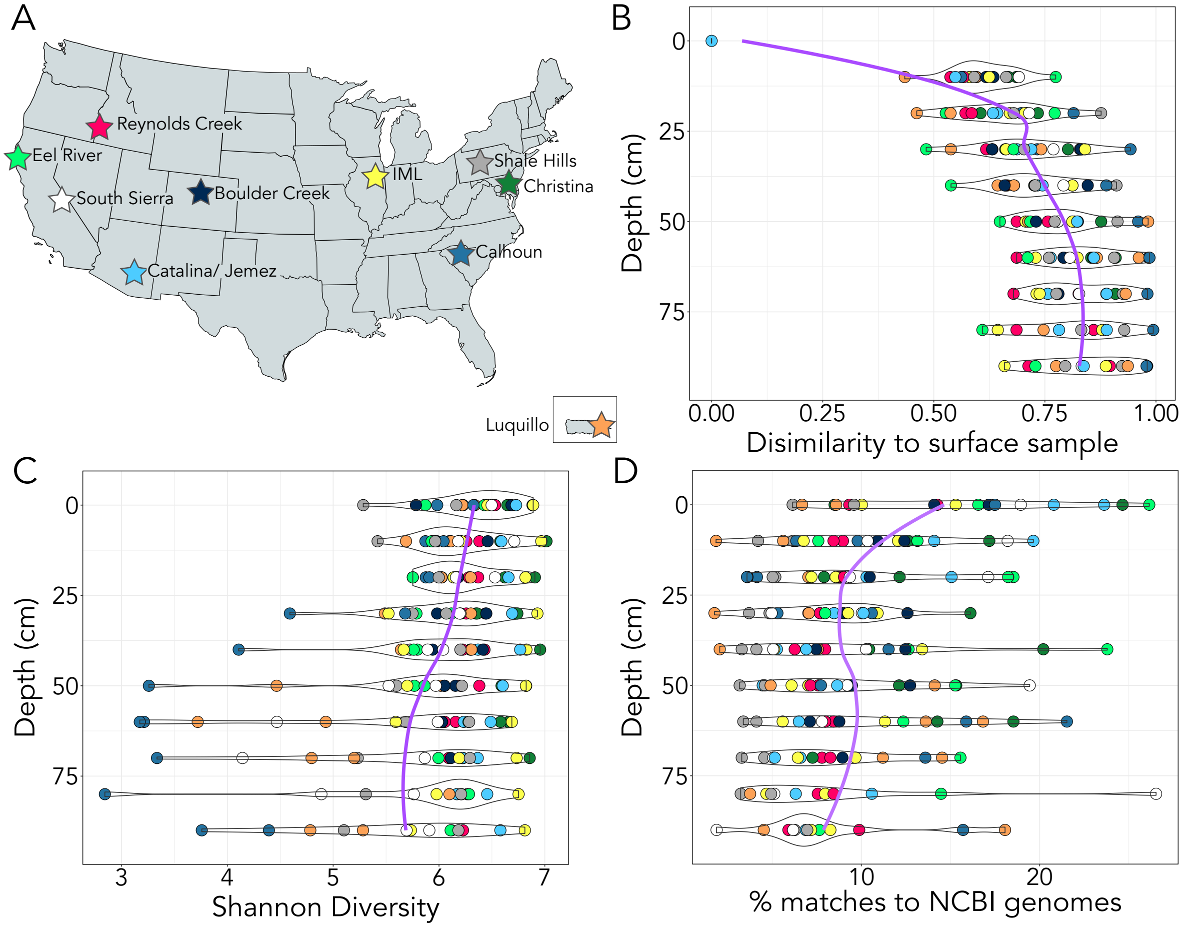


Figure 2.1: A) Site map of sampling locations. We analyzed bacterial and archaeal communities from 2 soil pits located at each of 10 different CZOs across the U.S. Each pit was sampled in 10 cm intervals from surface soils to one meter in depth (when possible). **B**) Bray-Curtis dissimilarity to surface samples increases with depth. As depth increases, soil bacterial and archaeal communities become less similar to those communities at the surface. **C**) Bacterial and archaeal diversity decreases with depth. Colors of points match the colors of the CZO sites indicated in panel A with two profiles sampled per site (n=20). **D**) The proportion of 16S rRNA gene

sequences from the sampled soils for which representative genome data are available decreases with depth. We matched our 16S rRNA gene amplicon sequences to 16S rRNA genes from finished bacterial and archaeal genomes in the NCBI database. At deeper soil depths, we found that fewer taxa in our dataset had representative genomes, indicating that the bacterial and archaeal taxa found in deeper soil horizons are less represented in genomic databases than those found in surface soils.

Some soil properties changed consistently with depth across all 20 profiles. Total N and organic C concentrations were both negatively correlated with soil depth, in agreement with previous observations (Jobbágy and Jackson, 2000; Marty et al., 2017) (depth vs. %C rho = -0.61, p<0.001; depth vs. %N rho = -0.56, p<0.001; Spearman). On average, soil total organic C concentrations below 50 cm were 4.4 times lower than in surface soils, while total N concentrations were 6.3 times lower. While we measured a suite of additional chemical and soil properties, only clay concentrations exhibited consistent changes with depth (with percent clay generally increasing with depth; rho = 0.29, p < 0.001; Spearman). Given that our sampling effort included a wide range of different soil types and the expectedly high degree of variability in inter- and intra-profile edaphic characteristics, our goal was not to determine if distinct soil samples harbored distinct microbial communities or to characterize the factors related to shifts in overall community composition. Rather, our goal was to determine if there were any consistent changes in soil microbial communities with depth across the 20 sampled profiles.

Community characteristics linked to soil depth

Unsurprisingly, we found that the location of each soil profile had a strong influence on the composition of soil bacterial and archaeal communities as determined by 16S rRNA gene amplicon sequencing (r = 0.47, p < 0.001, Permanova). Individual soil profiles generally harbored distinct microbial communities (Figure 2.2, Supplemental Figure S2.1). In addition to this variation across the profiles, soil depth also had a significant effect on the composition of the bacterial and archaeal communities within individual profiles (p < 0.01 for 16 of 20 profiles, rho values ranging from 0.24 - 0.45). In general, the variation in community composition with depth within a given profile, while significant, was typically less than the differences in soil communities observed across different profiles when all profiles and soil depths were examined together (Depth: r = 0.02, p < 0.001, Location: r = 0.47, p < 0.001, Permanova).

Several characteristics of the bacterial and archaeal communities changed consistently with depth despite the high degree of heterogeneity observed across the different soil profiles. As soil depth increased, microbial communities found at depth became increasingly dissimilar to those found in surface horizons (Figure 2.1B). When we analyzed the entire sample set together, dissimilarity to surface soils (0-10cm depth) was positively correlated with depth ($p < 0.001$, rho = 0.73, Spearman). This trend also held for 17 out of 20 individual soil profiles (depth was not significant in both Eel sites and IML site 1). We also found that the diversity of microbial communities (taxon richness) generally decreased with depth, with several CZOs exhibiting especially stronger declines with depth (Calhoun, Luquillo, and Southern Sierra) than others (Figure 2.1C). Lastly, when we compared the 16S rRNA gene sequences from this study to those 16S rRNA gene sequences from finished bacterial and archaeal genomes in the NCBI database, we found that the proportion of taxa for which genomic data is available declined with depth (from 6.2 - 26.1% in surface soils, to 1.9 - 18.0% in the deepest horizons sampled, Figure 2.1D). Although representative genomes are unavailable for the majority of soil bacterial and archaeal taxa (Lloyd et al., 2018), genomic information from closely-related taxa is available for a smaller proportion of taxa living at depth than those found in surface soil horizons.

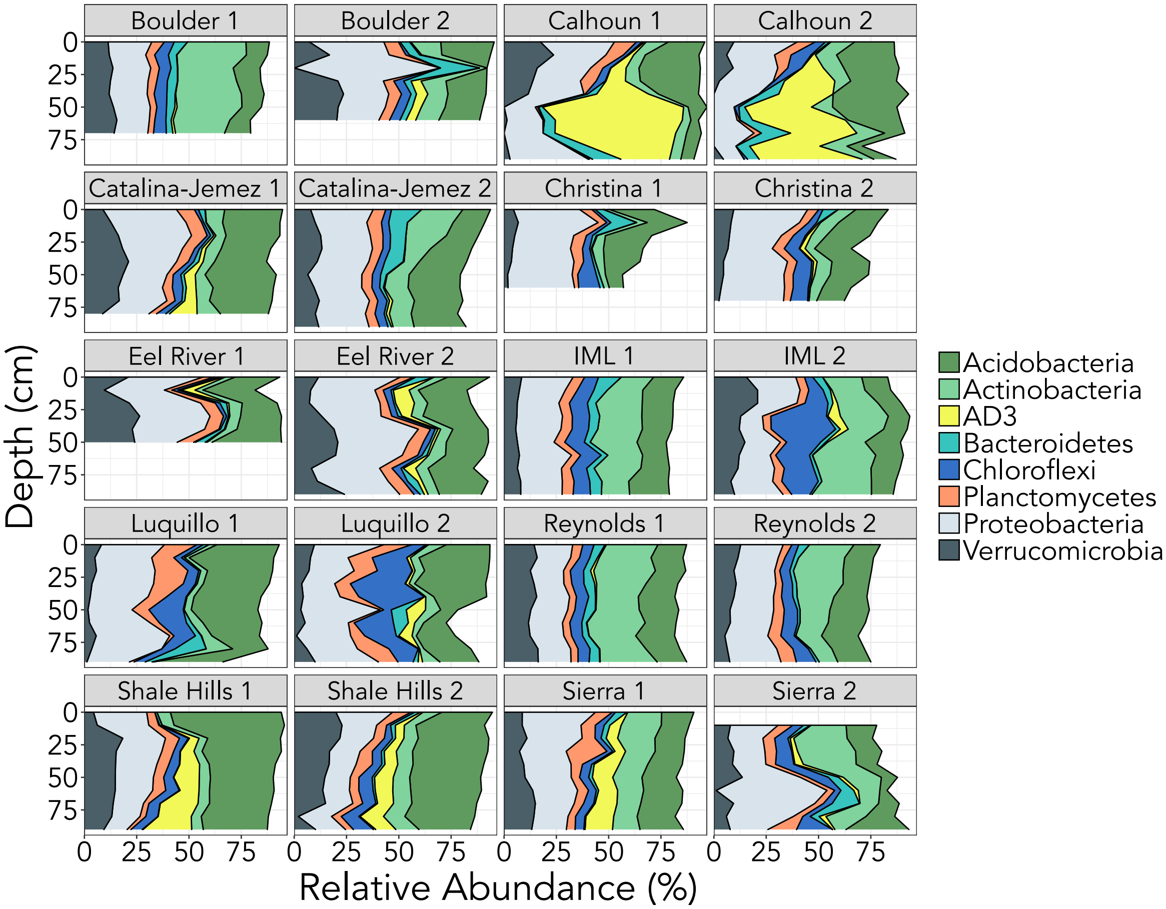


Figure 2.2: Different soil profiles have distinct microbial communities. Here we show the relative abundances of the eight most abundant phyla identified from our 16S rRNA gene amplicon sequencing effort. Not all profiles were sampled to one meter due to variable bedrock depth. Note that the two profiles sampled from each CZO site were selected to represent distinct soil types.

Taxonomic shifts with soil depth

Although each soil profile harbored distinct microbial communities (Figure 2.2), we identified five phyla that consistently increased in abundance with soil depth as measured by Spearman correlations across the entire dataset: Chloroflexi, Euryarchaeota, Nitrospirae, and the candidate phyla AD3 and GAL15 (Figure 2.3). For example, GAL15 and AD3 were typically 30 and 27 times more abundant in soils at 90 cm than in surface horizons, respectively. The candidate phylum AD3, Chloroflexi, and Nitrospirae have previously been found to increase in abundance with increasing soil depth in individual profiles (Will et al., 2010; Tas et al., 2014), while candidate phylum GAL15 has been shown to be abundant in oxic subsurface sediments (Lin et al., 2011). Members of these phyla are likely oligotrophic taxa adapted to survive in the resource-limited conditions found in deeper horizons. Indeed, soil Euryarchaeota (Leff et al., 2015), Chloroflexi, and Nitrospirae (Fierer et al., 2011) have been shown to decrease in abundance upon soil fertilization. These five phyla are also underrepresented in public genome databases; together, they account for only 2.8% of bacterial and archaeal genomes deposited in IMG (as of Dec 2018), reinforcing the observation highlighted in Figure 2.1D that poorly described taxa tend to be relatively more abundant in deeper soil horizons.

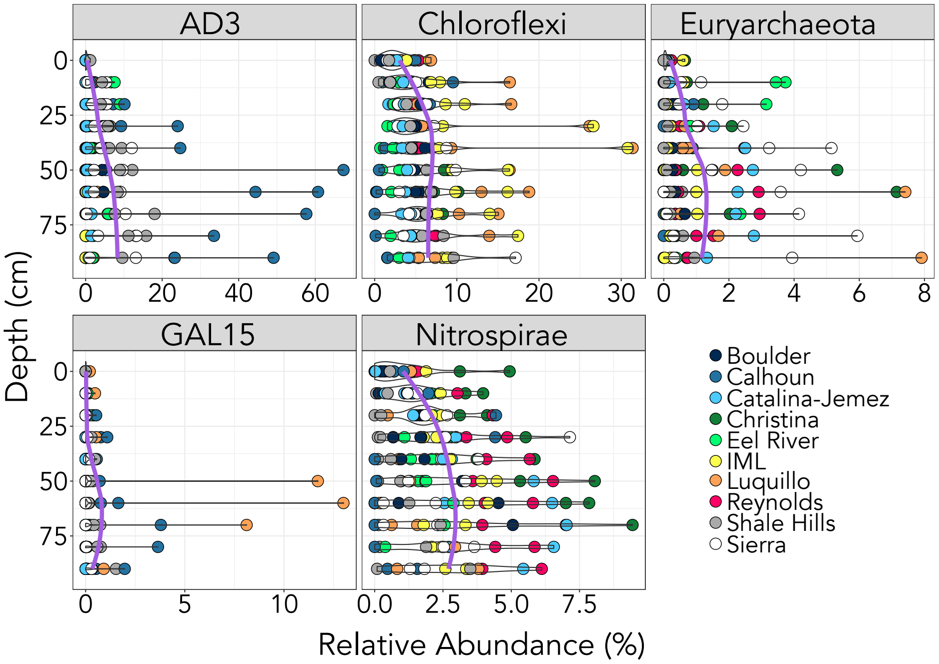


Figure 2.3: Five bacterial and archaeal phyla that consistently increased in relative abundance with soil depth. These phyla were identified via Spearman rank correlations against depth (FDR corrected p values < 0.02, rho > 0.22).

Community-level shotgun metagenomic analyses

We selected one soil profile from nine out of the original 10 CZO sites for shotgun metagenomic sequencing, targeting those profiles that displayed the most dissimilarity among different depths. Together, we obtained shotgun metagenomic data from 67 soil samples with an average of 7.84 million quality-filtered reads per sample. We first used these metagenomic data to quantify changes in the relative abundances of the bacterial, archaeal, and eukaryotic domains with depth. The overwhelming majority of rRNA gene sequences that we detected were from bacteria (89.2% - 98.7% of reads), followed by archaea (0.03% - 7.70%), and then eukaryotes

(0.04% - 4.27%). Interestingly, we found that the proportion of eukaryotic sequences in our samples decreased with depth (rho = -0.32, p = 0.05). Most of these eukaryotic rRNA gene reads were classified as Fungi (58%), then Charophyta (16%), Metazoa (9.3%), and Cercozoa (7.0%). These results are in line with previous work showing that the contributions of eukaryotes, most notably fungi, to microbial biomass pools typically decrease with soil depth (Turner et al., 2017).

We also directly compared the results obtained from our 16S rRNA amplicon and shotgun metagenomic sequencing across the same set of samples. We did this to check whether our PCR primers introduced significant biases in the estimation of taxon relative abundances. We found that the shotgun and amplicon-based estimations of the abundances of each of the eight phyla that were the most ubiquitous and abundant across the sampled profiles (Figure 2.2) were well correlated (Supplemental Figure S2.2, mean rho values = 0.70). Next, we checked whether our primers missed any major groups of bacteria or archaea, as it has been noted that many taxa from the Candidate Phyla Radiation (CPR, recently assigned to the superphylum Patescibacteria; Parks et al., 2018) are not detectable with the primer set used here (Eloe-Fadrosh et al., 2016). While we found that our primer pair did fail to recover sequences from the superphylum Patescibacteria, these taxa were rare in our data - the entire superphylum accounted for only 0.5% of 16S rRNA gene reads across the whole metagenomic dataset.

Candidate phylum AD3 is negatively correlated with organic carbon

We found that members of phylum AD3 were consistently more abundant in deeper soil horizons and particularly abundant in subsurface horizons from the Calhoun and Shale Hills CZOs (Figure 2.4). In these soils, AD3 dominated the microbial communities – in some samples, over 60% of 16S rRNA sequences were classified as belonging to members of the AD3 candidate phylum. The high abundances of AD3 were confirmed with shotgun metagenomic analyses (Supplemental Figure S2.2), indicating the abundances of this phylum were not inflated by PCR primer biases. Candidate phylum AD3 was first observed in a sandy, highly weathered soil from Virginia, U.S. (Zhou et al., 2003) and does not yet have a representative cultured isolate. Recently, the phylum was renamed Dormibacteraeota after three genomes were assembled from Antarctic soils (Ji et al., 2017). Other representative genomes from this phylum have also become available with the recent addition of 47 genomes assembled from thawing permafrost (Woodcroft et al., 2018). However, we refer to this phylum as 'AD3' to maintain consistency with other, previously published studies. The phylum AD3 has been observed in subsurface soil horizons previously (Kim et al., 2014; Billings et al., 2018), and its relative abundance has been found to be negatively correlated with water content, C, N, and total potential enzyme activities (Tas et al., 2014).

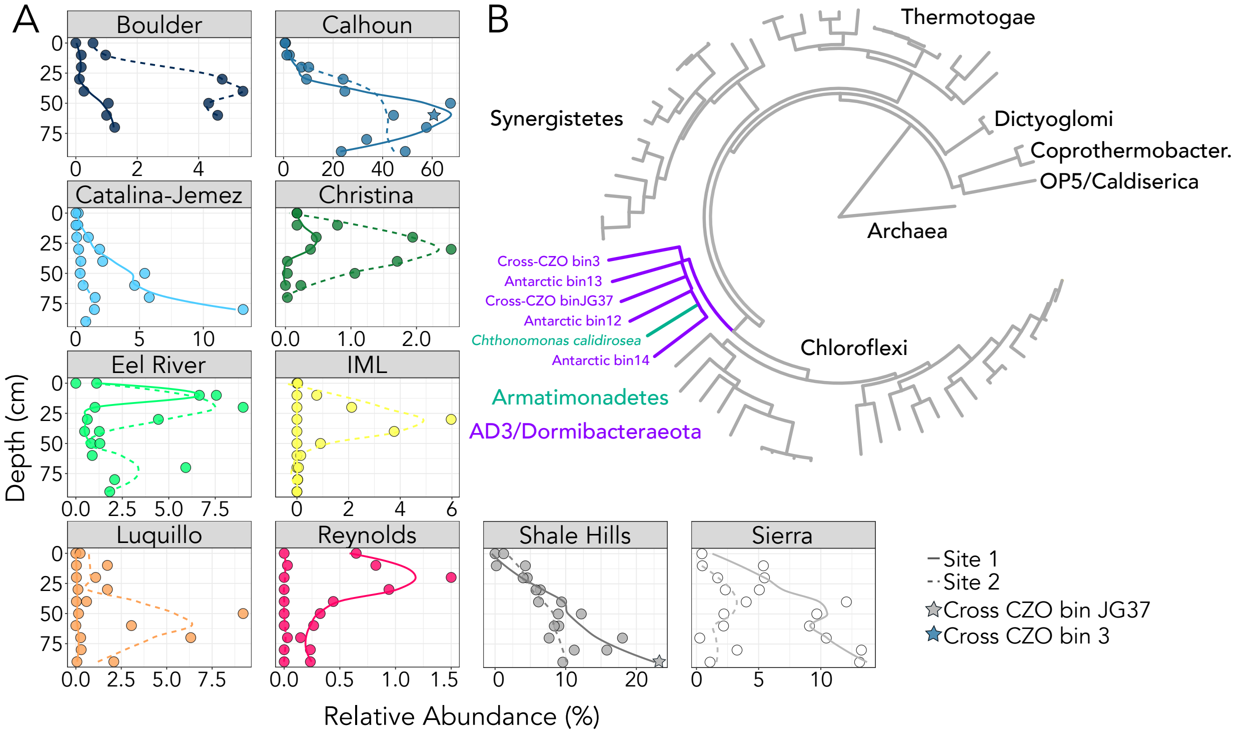


Figure 2.4: **A**) The relative abundance of phylum AD3 is variable across different soil profiles, but generally increases with depth. The samples used for the AD3 genome assemblies are noted with stars. **B**) The two AD3 genomes we assembled from the soil profile metagenomic data cluster phylogenetically with previously published AD3 genomes (Ji et al., 2017). Our deep soil AD3 genomes also fall between the known sister phyla Chloroflexi and Armatimonadetes, validating their identity as members of candidate phylum AD3. This tree was created using the concatenated marker gene phylogeny generated from checkM (Parks et al., 2015), and was plotted using ggtree (Yu et al., 2016) Only closely related phyla are included in the tree.

While the abundance of phylum AD3 was generally positively correlated with depth across all samples included in this study (rho = 0.22, p = 0.02, Spearman), this pattern did not hold for all profiles (Figure 2.4). Instead, we found organic C concentrations to be the best predictor of the abundance of AD3 in these soil communities (Supplemental Figure S2.3); AD3 was typically eight times more abundant in soils with less than 1% organic C than in soils where organic C concentrations were greater than 2%. Because soil depth and organic C concentrations were correlated across the profiles studied here, we used an independent dataset of surface soils (0-10 cm) collected from 1006 sites across Australia to determine if the abundances of AD3 were also correlated with organic C concentrations when analyses were restricted to a broad range of distinct surface soils (Bissett et al., 2016). Indeed, we found that the relative abundances of AD3 in the Australian surface soil dataset (which ranged from 0.0 to 7.0% of 16S rRNA gene sequences) were also negatively correlated with soil organic carbon concentrations (Supplemental Figure S2.3). Together these results indicate that AD3 is typically most abundant in surface or subsurface soils where organic C concentrations are relatively low. Additionally, given the high abundance of AD3 in many of the Australian surface soils and given that subsoil oxygen concentrations can remain relatively high (Hall et al., 2016), it is unlikely that soil-dwelling members of this phylum are obligate anaerobes.

AD3 draft genomes recovered from metagenomic data

To gain more insight into the potential traits and genomic attributes of soil AD3, we conducted deeper shotgun metagenomic sequencing on several soils where AD3 was found to be particularly abundant (Figure 2.4) with the goal of assembling draft genomes from members of this group. We were able to assemble two AD3 genomes, both from deep soils (Figure 2.4). These genomes are considered "substantially" complete according to checkM guidelines (Parks et al., 2015); bin 3 is estimated to be 72.7% complete at 3.4 Mb, while bin JG-37 is 74.54% complete at 3.0 Mb (further genome details in Supplemental Table S2.1). These genomes share only 47.4% average amino acid identity (AAI) (Konstantinidis and Tiedje, 2005) and cluster phylogenetically with the AD3 genomes assembled from Antarctic soil metagenomes (Ji et al., 2017), falling between the phyla Armatimonadetes and Chloroflexi (Figure 2.4).

Analyses of the AD3 genomes that we recovered indicate that members of this phylum are aerobic heterotrophs adapted to nutrient poor conditions. Both AD3 genomes encode high-affinity terminal oxidases, indicative of an aerobic metabolism (cbb$_3$ binJG37, bd bin3). These genomes contain no markers of an autotrophic metabolism, with no RuBisCO or hydrogenase genes detected in either of the assembled genomes. Both AD3 genomes contain trehalose 6-phosphate synthase, a key gene in the pathway for trehalose synthesis, a C storage compound that also confers resistance to osmotic stress and heat shock (Fung et al., 2013) and protects cells from oxidative damage, freezing, thermal injury, or desiccation stress (Kandror et al., 2002).

Additionally, both genomes contain glycogen catalysis (alpha-amylase, glucoamylases) and synthesis (glycogen synthase) genes. The ability to synthesize, store, and break down glycogen has been shown to promote the survival of bacteria during periods of starvation (Wilson et al., 2010; Fung et al., 2013). These attributes likely confer an advantage in resource-limited soils, as the ability to store C for later use may be advantageous in environments where organic C is infrequently available or of low quality.

Based on several lines of evidence, soil-dwelling AD3 appear to be oligotrophic taxa with low maximum growth rates. First, as mentioned above, these taxa have the highest relative abundances in soils with low organic C concentrations where we would expect oligotrophic lifestyles to be advantageous. Second, both AD3 genomes appear to encode a single rRNA operon, a feature often linked to low maximum potential growth rates (Roller et al., 2016). Third, although we cannot directly measure the maximum growth rate of uncultivated bacterial cells, we can estimate maximum growth rate from genomes by measuring codon usage bias with the ΔENC' metric (Novembre, 2002). ΔENC' is a measure of codon bias in highly expressed genes, and has been shown to correlate strongly with growth rate for both bacteria and archaea (Vieira-Silva and Rocha, 2009). We calculated ΔENC' for our AD3 genomes, the Antarctic AD3 genomes (Ji et al., 2017), the thawing permafrost AD3 genomes (Woodcroft et al., 2018), and a set of bacterial and archaeal genomes which matched the 16S rRNA gene amplicon sequences recovered from the soil profile samples at ≥99% sequence similarity. The ΔENC' values for all the AD3 genomes clustered together towards the lower end of the spectrum for our set of soil bacteria and archaea, indicating that members of the phylum AD3 are likely to exhibit low potential growth rates (Supplemental Figure S2.4).

To our knowledge, all previous AD3 genomes were recovered from either Antarctic desert (Ji et al., 2017) or permafrost soils (Woodcroft et al., 2018), while our genomes hail from subsurface soils collected from temperate regions. Despite these disparate origins, some central characteristics of the phylum AD3 appear to be consistent. Similar to the Antarctic AD3 genomes, our AD3 genomes also contained carbon-monoxide (CO) dehydrogenase genes. However, there are two types of CO dehydrogenases, which differ in their ability to oxidize CO and the rate at which they do so (King and Weber, 2007). While the active site of form I is specific to CO dehydrogenases, form II active sites also occur in many molybdenum hydroxylases that do not accept CO as a substrate (King and Weber, 2007). Using sequence data from our assembled AD3 genomes, the Antarctic AD3 genomes, and selected CO dehydrogenase large subunit sequences (coxL), we generated a phylogenetic tree based on the amino acid sequence of coxL (Supplemental Figure S2.5). With these analyses, we found that both of the AD3 genomes recovered here possess form II CO dehydrogenases, as do two of the Antarctic AD3 genomes. Although it has been shown that form II CO dehydrogenases can permit growth with CO as a sole C and energy source in some cases (Lorite et al., 2000), further work is needed to determine whether these genes allow AD3 to actively oxidize CO or if these genes code for molybdenum-containing hydroxylases responsible for other metabolic processes (Hille, 2005). Interestingly, one Antarctic AD3 genome also encodes a form I coxL, indicating that some members of this phylum are capable of CO oxidation (Supplemental Figure S2.5).

Analyses of our assembled AD3 genomes also reveal that these soil bacteria may be capable of spore formation. Altogether, our AD3 genomes contain 33 spore-related genes scattered across a variety of spore generation phases (Supplemental Table S2.2). Nutrient limiting conditions are known to trigger spore formation (Fujita and Losick, 2005), and sporulation can allow bacterial cells to persist until environmental conditions become more favorable. Additionally, members of the Chloroflexi, a sister phylum to AD3, are capable of spore formation (Cavaletti et al., 2006). Because there are no AD3 isolates available to test for sporulation, we adapted a method previously used in stool samples (Browne et al., 2016) to identify spore-forming taxa using a culture-independent approach. We incubated three soil samples from our study in 70% ethanol to kill all vegetative cells, and then used propidium monoazide (PMA) to block the amplification of DNA from these dead cells (Carini et al., 2016). We then sequenced these soils using our standard 16S rRNA gene amplicon method both with and without the ethanol and PMA treatment. We found that the abundances of the two dominant AD3 phylotypes were significantly higher in the spore-selected treatment than the untreated controls (Supplemental Table S2.3). Other known spore formers were enriched in the spore selection treatment as well, including taxa from the orders Actinomycetales, Bacillales (Browne et al., 2016), Myxococcales (Shimkets, 1999), and Thermogemmatisporales (Yabe et al., 2011).

Conclusions

Our results indicate that, as soil depth increases, not only do bacterial and archaeal communities become less diverse and change in composition, but novel, understudied taxa become proportionally more abundant in deeper soil horizons. We identified five poorly studied bacterial and archaeal phyla that become more abundant in deeper soils across a broad range of locations, and investigated one of these further (the candidate phylum AD3) to determine what characteristics may allow AD3 to survive and dominate in resource-limited soil environments. We found that members of AD3 are likely slow-growing aerobic heterotrophs capable of persisting in low resource conditions by putatively storing and processing glycogen and trehalose. Members of this candidate phylum also contain type I and II carbon monoxide dehydrogenases, which can potentially enable the use of trace amounts of CO as a supplemental energy source. We also found that soil-dwelling AD3 are likely capable of sporulation, another trait that may allow cells to persist during periods of limited resource availability. More generally, analyses of these novel members of understudied phyla suggest life history strategies and traits that may be employed by oligotrophic microbes to thrive under resource-limited soil conditions.

CHAPTER III

GENOME REDUCTION IN AN ABUNDANT AND UBIQUITOUS SOIL BACTERIUM ‘CANDIDATUS UDAEOBACTER COPIOSUS’

Adapted from:
Brewer TE, Handley KM, Carini P, Gilbert JA, Fierer N. (2016). Genome reduction in an abundant and ubiquitous soil bacterium ‘*Candidatus* Udaeobacter copiosus’. *Nature Microbiology* **2**: 16198.

Abstract

Although bacteria within the Verrucomicrobia phylum are pervasive in soils around the world, they are underrepresented in both isolate collections and genomic databases. Here we describe a single verrucomicrobial group within the class Spartobacteria that is not closely related to any previously described taxa. We examined >1000 soils and found this spartobacterial phylotype to be ubiquitous and consistently one of the most abundant soil bacterial phylotypes, particularly in grasslands, where it was typically the most abundant. We reconstructed a nearly complete genome of this phylotype from a soil metagenome for which we propose the provisional name ‘*Candidatus* Udaeobacter copiosus’. The *Ca.* U. copiosus genome is unusually small for a cosmopolitan soil bacterium, estimated by one measure to be only 2.81 Mbp, compared to the predicted effective mean genome size of 4.74 Mbp for soil bacteria. Metabolic reconstruction suggests that *Ca.* U. copiosus is an aerobic heterotroph with numerous putative amino acid and vitamin auxotrophies. The large population size, relatively small genome and multiple putative auxotrophies characteristic of *Ca.* U. copiosus suggest that it may be undergoing streamlining selection to minimize cellular architecture, a phenomenon previously thought to be restricted to aquatic bacteria. Although many soil bacteria need relatively large, complex genomes to be successful in soil, *Ca.* U. copiosus appears to use an alternate strategy, sacrificing metabolic versatility for efficiency to become dominant in the soil environment.

Introduction

Soils harbor massive amounts of undescribed microbial diversity. For example, more than 120,000 unique bacterial and archaeal taxa were found in surface soils of Central Park in New York City, of which only ~15% had 16S rRNA gene sequences matching those contained in reference databases and <1% had representative genome sequence information (Ramirez et al., 2014). This undescribed soil microbial diversity is not evenly distributed across the tree of life. For example, Acidobacteria and Verrucomicrobia, two of the more abundant bacterial phyla found in soil (Janssen, 2006; Bergmann et al., 2011) represent only 0.08% and 0.06% of all cultured bacterial isolates in the Ribosomal Database Project (RDP) (Wang et al., 2007) and only 0.08% and 0.14% of publicly-available bacterial genomes found in Integrated Microbial Genomes (IMG; Chen et al., 2017), respectively. Although the ecology and genomic attributes of abundant soil taxa are beginning to be described (VanInsberghe et al., 2015), we still lack basic information on the vast majority of soil microbes. These knowledge gaps highlight that a huge fraction of living biomass in terrestrial systems remains enigmatic (Fierer et al., 2009) and that we are only beginning to identify the influence of specific microbes on soil biogeochemistry and fertility.

For this study, we focus our exploration of undescribed microbial diversity on the Verrucomicrobia phylum. Although Verrucomicrobia are generally recognized as being among the most numerically abundant taxa in soil (Janssen, 2006; Bergmann et al., 2011) we know very little about the ecological or genomic attributes that contribute to their success. The phylum Verrucomicrobia is highly diverse and its members possess a broad range of metabolic capabilities. For example, members of the class Methylacidiphilae are nitrogen-fixing acidophiles capable of methane oxidation (Dunfield et al., 2007) while *Akkermansia muciniphila* of the class Verrucomicrobiae is a mucin-degrading resident of the human gut (Everard et al., 2013). However, the dominant Verrucomicrobia found in soil typically belong to the class Spartobacteria. While Verrucomicrobia accounted for >50% of all bacterial 16S rRNA gene sequences in tallgrass prairie soils in the United States, >75% of these sequences were assigned to the class Spartobacteria (Fierer et al., 2013). Currently, the class Spartobacteria contains only a single described and sequenced isolate, *Chthoniobacter flavus*, a slow-growing aerobic heterotroph capable of using common components of plant biomass for growth (Sangwan et al., 2004; Kant et al., 2011). While Spartobacteria are prevalent in soils, they have also been observed in marine systems (*Spartobacteria baltica*; Herlemann et al., 2013) and as nematode symbionts (genus *Xiphinematobacter*; Vandekerckhove et al., 2000).

Here we report the distribution of a dominant Spartobacteria lineage, compiling data from both amplicon and shotgun metagenomic 16S rRNA gene surveys to quantify its relative abundance across >1000 unique soils. We assembled a near-complete genome of this lineage from a single soil where it was exceptionally abundant. These results provide our first glimpse into the phylogeny, ecology, and potential physiological traits of a dominant soil Verrucomicrobia and suggest that members of this group are efficient at growing and persisting in the low resource conditions common in many soil microenvironments.

Results and Discussion

Distribution of the dominant Verrucomicrobia in soil

A single spartobacterial clade dominates bacterial communities found in a wide range of soil types across the globe. One phylotype from this group of Spartobacteria represented up to 31% of total 16S rRNA gene sequences recovered from prairie soils (Fierer et al., 2013). This phylotype shares 99% 16S rRNA gene sequence identity with a ribosomal clone named ‘DA101’, first described in 1998 as a particularly abundant 16S rRNA sequence recovered from grassland soils in the Netherlands (Felske and Akkermans, 1998). To determine if the DA101 phylotype (termed ‘DA101’ herein) is abundant in other soils, we re-analyzed amplicon 16S rRNA gene sequence data obtained from >1000 soils representing a wide range of soil and site characteristics. We found that DA101 was on average ranked within the top two most abundant bacterial phylotypes in each study (Figure 3.1). In over 70% of the soils analyzed DA101 was within the top ten most abundant phylotypes. Interestingly, other phylotypes belonging to the same family as DA101 (Chthoniobacteraceae) were also found within the top 5 most abundant phylotypes of several studies (Figure 3.1).

As some 16S rRNA gene PCR primer sets can misestimate the relative abundance of Verrucomicrobia (Guo et al., 2016; Bergmann et al., 2011), we investigated whether the apparent numerical dominance of DA101 in amplicon datasets was a product of PCR primer biases. To do so, we quantified the abundance of DA101 16S rRNA genes within previously published soil shotgun metagenomes (Leff et al., 2015; Fierer et al., 2012). The relative abundance of DA101 in amplicon data was well correlated with the relative abundance of

DA101 in shotgun metagenomic data ($P < 0.0001$, rho = 0.50, n = 102). Confirming the amplicon-based results (Figure 3.1), we found that DA101 was also among the most abundant phylotypes observed in the soil bacterial communities characterized via shotgun metagenomic sequencing (Supplementary Figure S3.1). Therefore, we conclude that the numerical dominance of DA101 in soils is not simply a product of primer biases.

Despite DA101 being one of the most abundant phylotypes found in soil, its proportional abundance can vary significantly across soil types (Figure 3.1 and Supplementary Figure S3.1). We used metadata associated with each soil sample to determine which of the measured soil and site characteristics best predicted the relative abundance of DA101. We found that DA101 was significantly more abundant in grassland soils than in forest soils ($P < 0.0001$, n = 64, Mann-Whitney test, Supplementary Figure S3.2); on average, DA101 is six times more abundant in grassland soils. These findings indicate that the soils in which DA101 excels do not overlap with those forest soils dominated by non-symbiotic Bradyrhizobium taxa, another ubiquitous and abundant group of soil bacteria (VanInsberghe et al., 2015). Across the grassland soils included in our meta-analysis, the relative abundance of DA101 was positively correlated with both soil microbial biomass ($P < 0.0001$, rho = 0.57, n = 31, Spearman, Supplementary Figure S3.3), and aboveground plant biomass ($P < 0.0001$, rho = 0.47, n = 366, Spearman, Supplementary Figure S3.3). Together, these results suggest that DA101 prefers soils receiving elevated amounts of labile carbon inputs. We did not identify any consistently significant correlations between the abundance of DA101 and other prokaryotic or eukaryotic taxa, suggesting that DA101 is unlikely to be a part of an obligate pathogenic or symbiotic relationship.

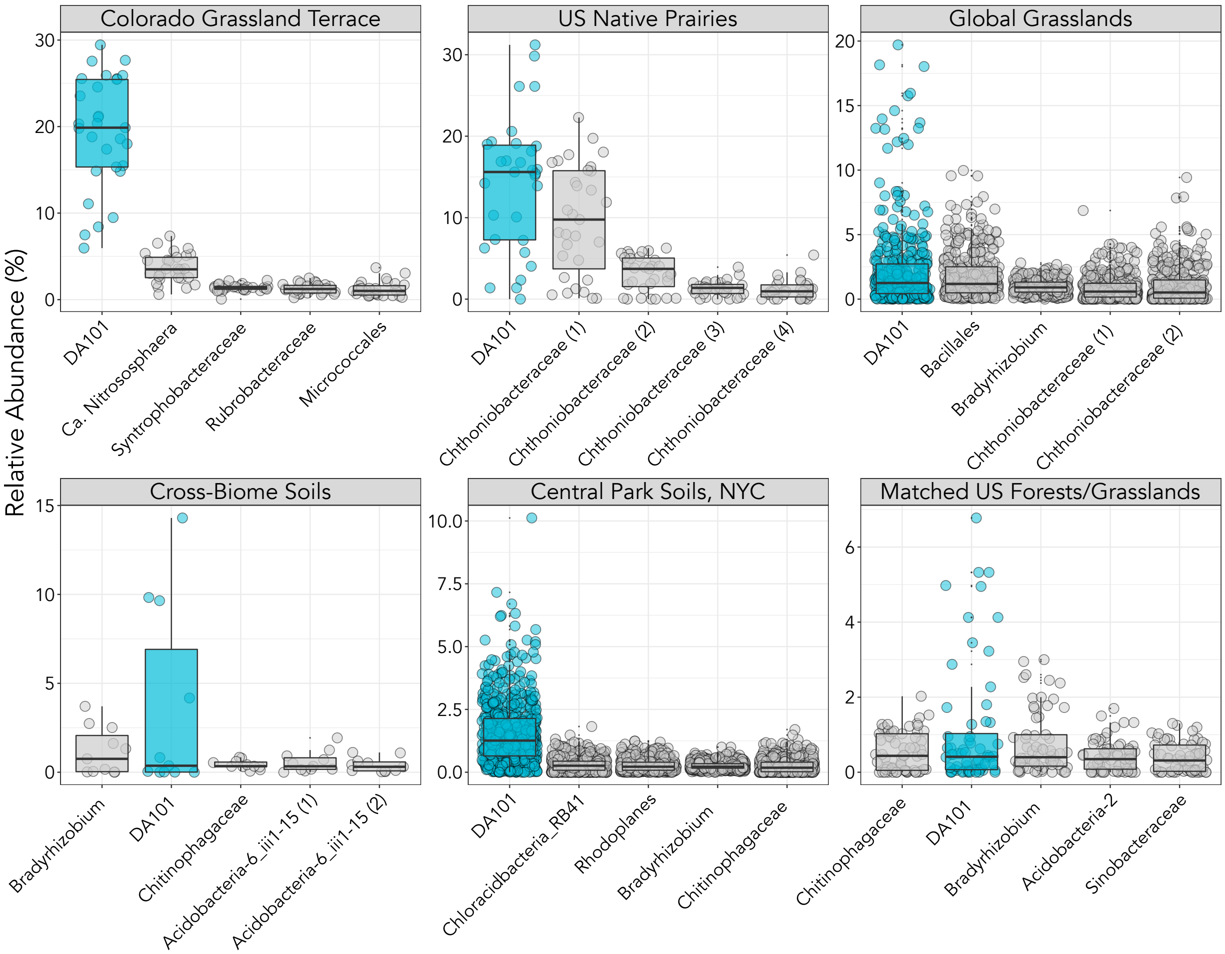


Figure 3.1: DA101 is one of the most abundant bacterial phylotypes found across >1000 soils collected from a wide range of ecosystem types throughout the world. The DA101 phylotype is indicated in blue while other abundant taxa are indicated in grey. Taxa are listed on the x-axis in order of their median rank abundance (taxa on the left are the most abundant). Stars denote data sets from previously published studies, from left to right: Fierer et al. 2013, Leff et al. 2015, Fierer et al. 2012, Ramirez et al. 2014, and Crowther et al. 2014.

Diversity of soil Verrucomicrobia

We determined the phylogenetic placement of DA101 and other soil Verrucomicrobia by assembling near full-length 16S rRNA gene sequences from six distinct grassland soils collected from multiple continents (Figure 3.2, Supplementary Table S3.1). Although we were able to assemble representative 16S rRNA gene sequences from all verrucomicrobial classes except Methylacidiphilae, 93% of verrucomicrobial sequences fell within the Spartobacteria class and 87% of these fell within the DA101 clade. These phylogenetic analyses confirm that DA101 belongs to the class Spartobacteria (Figure 3.2). However, within the Spartobacteria class, the

DA101 clade is clearly distinct from the clade containing *Chthoniobacter flavus* (Sangwan et al., 2004; Kant et al., 2011), as DA101 shares only 92% 16S rRNA gene sequence identity with *C. flavus*. These findings indicate that DA101 is likely a representative of a new verrucomicrobial genus. We propose the candidate genus name '*Candidatus* Udaeobacter' for the DA101 clade; the proposed name combines Udaeus ('of the earth', Greek) with bacter ('rod' or 'staff', Greek), and like Chthoniobacter refers to one of the Spartoi of the Cadmus myth. We recommend the provisional name '*Candidatus* Udaeobacter copiosus' for the DA101 phylotype, which refers to its numerical dominance in soil.

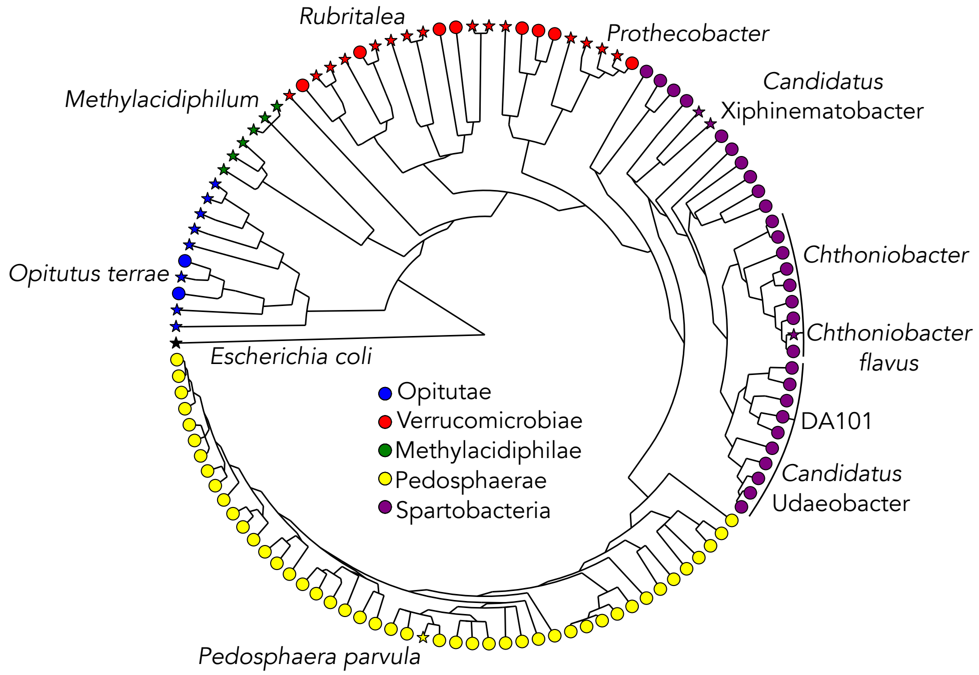


Figure 3.2: Phylogenetic analyses of soil Verrucomicrobia. Stars denote 16S rRNA gene sequences of named isolates while circles represent environmental 16S rRNA gene sequences assembled from 6 soils using EMIRGE (Miller et al., 2013) (Supplementary Table S3.1). The uncultivated verrucomicrobial phylotype DA101 falls within a cluster distinct from cultivated Spartobacteria. Notable verrucomicrobial isolates and genera are labeled. Colors indicate verrucomicrobial classes.

Draft genome of '*Ca.* Udaeobacter copiosus' recovered from metagenomic data

Despite their ubiquity and abundance in soil, there is no genomic data currently available for any representative of the '*Candidatus* Udaeobacter' clade. Typically, soil hyper-diversity confounds the assembly of genomes from metagenomes (Howe et al., 2014), requiring single-cell analysis or laboratory isolation to produce an assembled genome. However, we leveraged the sheer abundance of *Ca.* U. copiosus in an individual soil to obtain a nearly complete genome from metagenomic data. We deeply sequenced a soil where *Ca.* U. copiosus accounted for >30% of 16S rRNA gene sequences and assembled a draft genome from the resulting metagenome. We used GC content, coverage, tetranucleotide frequencies, and the phylogenetic affiliation of predicted proteins to bin assembled contigs, resulting in a draft *Ca.* U. copiosus genome with 238 contigs. The draft genome is 2.65 Mbp in size, has a GC content of 54%, and encodes for 3,042 predicted proteins, 67% of which could be assigned to Pfam protein families (Finn et al., 2016) by the IMG annotation pipeline (Chen et al., 2017).

The *Ca.* U. copiosus genome shares only 69.3% average nucleotide identity (Varghese et al., 2015) with the genome of its closest sequenced relative *C. flavus*, further supporting its proposed placement in the distinct genus '*Candidatus* Udaeobacter'. While no 16S rRNA gene was assembled within the *Ca.* U. copiosus genome, we used Metaxa2 (Bengtsson-Palme et al., 2015) to extract fragments of a single DA101-like 16S rRNA gene from the raw metagenomic sequences we used for assembly. This 16S rRNA gene has 100% identity to the DA101 amplicon sequence and has the same average coverage (23-29x) as the *Ca.* U. copiosus genome (27x), suggesting this genome belongs to a representative of the DA101 clade. As a second measure to verify this genome is a representative of the DA101 clade, we compared the abundance of three housekeeping genes assembled within the *Ca.* U. copiosus genome (dnaK, rpoB, and secY) to the abundance of the DA101 16S rRNA gene in >100 metagenomic samples from two separate studies (Leff et al., 2015; Fierer et al., 2012). All three genes show a very strong significant correlation with the DA101 16S rRNA gene ($P < 0.0001$, rho > 0.87, $n = 102$, Pearson correlation, Supplementary Figure S3.4), further evidence that this genome represents the DA101 clade and that this lineage is as abundant in soil as our analyses based on the 16S rRNA gene suggest.

We estimate that the full *Ca.* U. copiosus genome will be approximately 2.81 Mbp in length based on the recovery of 94% of domain-specific single copy housekeeping genes commonly used to estimate genome completion (Ciccarelli et al., 2006). Based on this estimate, *Ca.* U. copiosus appears to have a particularly small genome size compared to *C. flavus* and other sequenced heterotrophic soil Verrucomicrobia (Supplementary Table S3.2). Indeed, the genome size of *Ca.* U. copiosus is much more similar to Verrucomicrobia of the class Methylacidiphilae (Hou et al., 2008) - thermophiles for whom genome size and growth temperature are negatively correlated (Sabath et al., 2013). To determine how the genome size of *Ca.* U. copiosus compares to other soil bacteria, we compiled data from 378 finished and permanent draft genomes in IMG whose 16S rRNA gene sequences matched the 16S rRNA gene amplicon sequences obtained by Leff et al. (2015) with at least 99% identity. Nearly all of these 378 bacterial genomes were from cultivated taxa (99%). We estimated the genome completeness for each of the 378 taxa using the same domain specific marker genes as for *Ca.* U. copiosus and found the mean estimated genome size of these taxa to be 5.28 ± 2.15 Mbp (mean ± SD), which is similar to metagenomic based estimates of mean genome size for soil microbes (4.74 ± 0.69) (Raes et al., 2007). Strikingly, the estimated 2.81 Mbp genome of *Ca.* U. copiosus is ~50% smaller than the mean genome size of these 378 taxa; only 48 (13%) of these genomes are smaller than *Ca.* U. copiosus. Furthermore, the majority (65%) of soil taxa with genomes smaller than *Ca.* U. copiosus originate from organisms with obligate intracellular or host-associated lifestyles (Figure 3.3).

Although soil bacteria with larger genomes tend to be more common in soil, *Ca*. U. copiosus seems to be a notable exception to this pattern. We linked the genome size of each of the matched IMG bacterial genomes with the average abundance of their corresponding amplicon sequence from Leff et al. 2015 and found that genome size is positively correlated with average relative abundance ($P < 0.001$, rho = 0.37, n = 378, Spearman, Figure 3.3). That is, sequenced bacteria with large genomes tend to comprise a significantly larger proportion of soil bacterial communities. On average, the genomes of soil prokaryotes are larger than those inhabiting aquatic ecosystems (Giovannoni et al., 2014) or the human gut (Nayfach and Pollard, 2015). These relatively large genomes are thought to provide soil-dwelling bacteria with a more diverse genetic inventory to enhance survival in conditions where resources are diverse, but sparse (Konstantinidis and Tiedje, 2004; Barberán et al., 2014). However, the *Ca*. U. copiosus genome has a conspicuously reduced genome given its abundance (Figure 3.3). This suggests that *Ca*. U. copiosus occupies a niche space that does not require expansive functional diversity and points to an alternative route to success for soil bacteria. These results also suggest that abundant, uncultivated soil bacteria likely have smaller genomes than the cultivated taxa that represent the majority of available genomic data. A similar pattern has been observed in aquatic systems, where uncultivated taxa often have smaller genomes than cultivated taxa (Button and Robertson, 2001). Because most genomic information is derived from cultivated bacterial taxa, the lack of genomic information from bacteria with compact genomes may stem from challenges associated with culturing taxa with reduced genomes (Giovannoni et al., 2014).

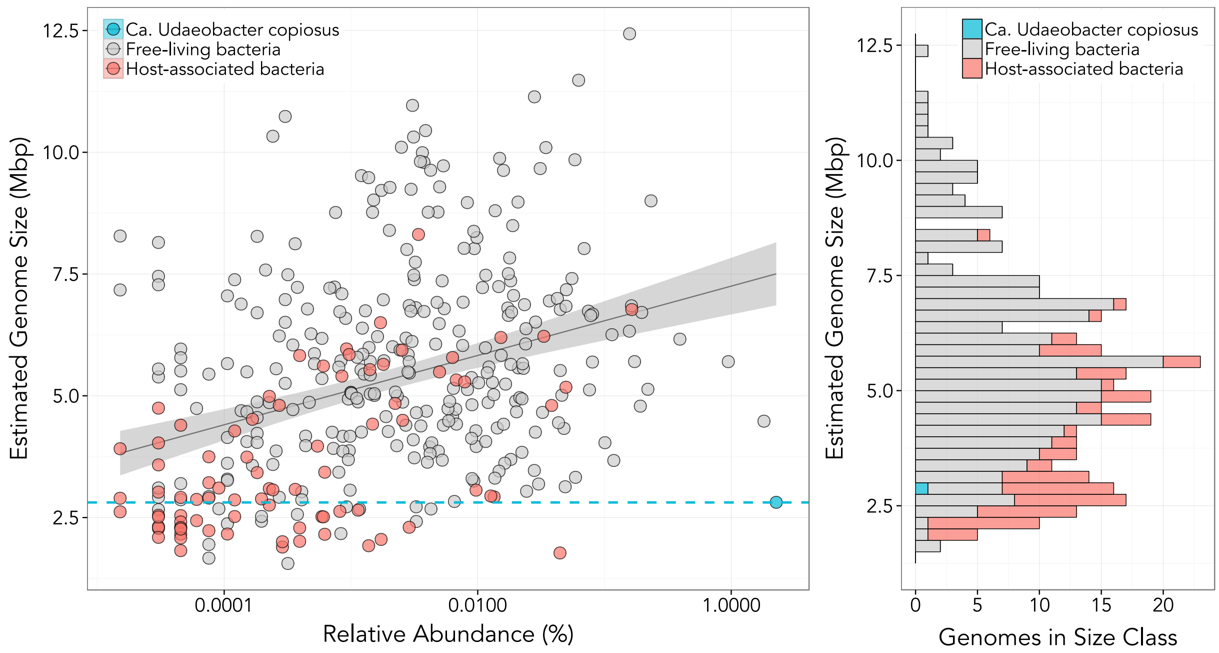


Figure 3.3: '*Ca.* Udaeobacter copiosus' has a reduced genome size compared to other abundant grassland soil bacteria. **A**) Points represent the estimated genome size and relative abundances of 378 bacterial genomes obtained by matching 16S rRNA gene sequences from Leff et al. 2015 to 16S rRNA gene sequences extracted from IMG genomes at 99% sequence identity. This dataset focused on surface soils collected from grasslands across the globe; the average abundances shown here may not apply to other soil or ecosystem types. Only genomes classified as 'permanent draft' or 'finished' status were used. Bacteria with larger genomes tend to be more abundant ($p < 0.0001$, rho = 0.368, n = 378, Spearman correlation), with *Ca.* U. copiosus (indicated in blue) being a notable exception to this pattern, as it has a high relative abundance (2.26% of 16S rRNA sequences) but a relatively small genome. The shaded region represents the 95% confidence interval of the trend line. **B**) Host-associated bacteria make up a majority of sequenced small genomes in soil. In the genome size range for *Ca.* U. copiosus (2.75 Mbp - 3.00 Mbp), 56% of soil taxa have a host-associated lifestyle.

Metabolic reconstruction of the *Ca.* U. copiosus genome points to an aerobic heterotrophic lifestyle with the capacity to use a limited range of carbon substrates for growth including glucose, pyruvate, and chitobiose. Glycogen/starch synthesis and utilization genes were identified (glgABCP and amyA), suggesting that *Ca.* U. copiosus has the capacity to store surplus carbon as glycogen or starch. Glycogen metabolism has been demonstrated in other Verrucomicrobia (Khadem et al., 2012). Genes encoding for the complete biosynthesis of vitamins B_2, B_3, B_5 (from valine) and B_6 were recovered, as well as full biosynthetic pathways for de novo synthesis of alanine, aspartate, asparginine, glutamate, glutamine, lysine, serine, and proline. Nearly complete pathways were recovered for glycine, threonine and methionine biosynthesis (Supplementary Figure S3.5). Genes encoding for the conversion of methionine to cysteine were present as the only apparent route to cysteine biosynthesis. Genes indicative of autotrophic metabolism (for example, RuBisCO, ATP citrate lyase) were not identified. Additionally, genes indicative of methanotrophy (pmo), methylotrophy (mxaF or xoxF), ammonia (amo) or nitrite oxidation (nxr) were not found.

Genes encoding for the biosynthesis of all branched-chain (isoleucine, leucine and valine) and aromatic (tryptophan, tyrosine and phenylalanine) amino acids were conspicuously underrepresented in the *Ca*. U. copiosus genome. The biosynthetic pathways for arginine and histidine were also incomplete (Supplementary Figure S3.5), along with the entire vitamin B_{12} synthesis pathway, despite the presence of three genes encoding vitamin B_{12}-dependent proteins (methionine synthase, ribonucleotide reductase, and methylmalonyl-CoA mutase). It is conceivable that genomic information encoding for these putative auxotrophies is present on genome fragments that were not recovered in our metagenome assembly, or is encoded on extrachromosomal elements that are commonly missed in metagenomic assemblies (for example, a plasmid; Jørgensen et al., 2015). Relative to *C. flavus*, 34 amino acid biosynthetic genes are needed for *Ca*. U. copiosus to be fully prototrophic. In *C. flavus*, these genes are not organized on operons (Kant et al., 2011), meaning they are likely randomly distributed throughout the *Ca*. U. copiosus genome as well. Moreover, the absence of branched-chain amino acid and histidine synthesis pathways in the *Ca*. U. copiosus genome is consistent with previous observations that branched chain and histidine biosynthesis genes are underrepresented in native prairie populations of soil Verrucomicrobia (Fierer et al., 2013). Additionally, plasmids are uncommon within isolates of the Verrucomicrobia phylum, with only one species known to maintain a plasmid - Opitutaceae Bacterium Strain TAV5 (Kotak et al., 2015), a distant relative to *Ca*. U. copiosus.

Auxotrophy in free-living bacteria is not expected to be a rare phenomenon; one study estimated that 85% of free-living bacteria have at least one vitamin or amino acid auxotrophy (D'Souza et al., 2014) and multiple studies have shown that auxotrophic mutants have a pronounced growth advantage over their wildtype counterparts when supplied with the compounds they can not synthesize (D'Souza et al., 2014; Kim and Levy, 2008). Vitamin B_{12}

auxotrophies are relatively common in soil (Lochhead, 1958), suggesting this metabolically expensive vitamin is generally available to many soil bacteria. Similarly, the eight amino acids that we did not identify complete pathways for in the *Ca*. U. copiosus genome are among the most energetically expensive to make (Akashi and Gojobori, 2002) (Supplementary Figure S3.5). This suggests that if *Ca*. U. copiosus is auxotrophic for some of these metabolites, acquiring them from the environment would provide *Ca*. U. copiosus an energetic savings relative to taxa that synthesize them de novo.

Although *Ca*. U. copiosus appears to lack genes for several amino acid synthesis pathways, numerous genes encoding for peptide transport, degradation and recycling were identified. Indeed, when scaled for genome size, *Ca*. U. copiosus encodes four times as many putative peptide and amino acid transporters as *C. flavus* (1.5% of genome to 0.37%) and twice as many predicted proteases (6.5% of genome versus 3.2%). *Ca*. U. copiosus also encodes for all components of the bacterial proteasome. Proteasomal degradation is critical for amino acid recycling under starvation conditions in mycobacteria (Elharar et al., 2014). The enrichment of peptide transport and degradation systems in the *Ca*. U. copiosus genome suggest that at least some of the amino acids *Ca*. U. copiosus appears incapable of synthesizing are available directly from the soil environment or by associations with other soil biota.

Ca. U. copiosus clearly has a reduced genome size compared to other soil bacteria (Figure 3.3) and other Verrucomicrobia with similar lifestyles (Supplementary Table S3.2). Bacterial genome reduction is thought to occur through two main mechanisms, genetic drift and streamlining selection, both mediated by extremes in effective population sizes (N_e; reviewed in Batut et al., 2014). The effect of genetic drift on microbes with a small N_e and low recombination rates, such as endosymbiotic bacteria, leads to the accumulation of deleterious mutations and subsequent loss of genetic material, which is typically identifiable in genomes by the presence of numerous pseudogenes and large noncoding intergenic regions (Batut et al., 2014). In contrast, free-living organisms with a large N_e are thought to undergo 'streamlining' selection to minimize genome size (Batut et al., 2014; Giovannoni et al., 2014). The genome-streamlining hypothesis proposes that, in large bacterial populations, reduced genome complexity is a trait under natural selection, especially in environments where nutrients can be sparse and periodically limit growth (Giovannoni et al., 2014).

The abundance, putative auxotrophies, and cosmopolitan distribution of *Ca.* U. copiosus (Figure 3.1), together with its small genome size relative to other soil microbes (Figure 3.3) and Verrucomicrobia with similar lifestyles (Supplementary Table S3.2), suggests that its small genome is a product of streamlining selection. Although it is difficult to accurately measure N_e in wild populations of bacteria, evidence of drift-mediated genome reduction was not present in the *Ca.* U. copiosus genome (such as large numbers of pseudogenes or unusually large intergenic spaces). Although all contemporary free-living organisms with streamlined genomes inhabit aquatic environments (Giovannoni et al., 2014; Kantor et al., 2013), compared to these aquatic environments, soil is more heterogeneous (Vos et al., 2013), has greater overall microbial diversity (Fierer and Lennon, 2011), and slower carbon turnover (Giovannoni and Vergin, 2012). Therefore, the functional complexity required by soil microbes to succeed within a given niche is likely large relative to that required by aquatic microbes. This means that the effects of genome streamlining are likely to be most evident (i.e., result in smaller genomes) in aquatic environments. This expectation is reflected in the fact that, on average, the genomes of aquatic microbes are smaller than their terrestrial counterparts (Button and Robertson, 2001). However, the small genome and numerous putative pathways missing from *Ca.* U. copiosus suggest that genome streamlining may not be unique to aquatic organisms and that genome streamlining may also confer a selective growth advantage in the soil environment.

The probable effects of genome streamlining in *Ca.* U. copiosus seem to have resulted in reduced catabolic and biosynthetic capacity, and thus an apparent loss of metabolic versatility. The underrepresentation of multiple costly amino acid and vitamin biosynthetic pathways in the *Ca.* U. copiosus genome implies that these compounds can be acquired from the soil environment. Several studies have shown that free amino acids and oligopeptides are present in soil (Friedel and Scheller, 2002; Farrell et al., 2013). The enrichment of proteases and amino acid and peptide importers in the *Ca.* U. copiosus genome suggests that it is well equipped to assimilate this fraction of soil organic matter. Dispensing the capacity to synthesize costly amino acids and vitamins would likely provide *Ca.* U. copiosus a growth advantage in resource limiting conditions when competition for labile carbon is high. Furthermore, many of the amino acids and vitamins *Ca.* U. copiosus appears unable to synthesize are involved in synergistic growth (Mee *et al.*, 2014) and may be supplied by other microbes as common community goods (Morris et al., 2012). Based on the few spartobacterial isolates that have been cultivated (Sangwan et al., 2004), culture-independent studies (Fierer et al., 2013; Portillo et al., 2013), and the genomic data presented here, we speculate that *Ca.* U. copiosus is a small, oligotrophic soil bacterium that reduces its requirement for soil organic carbon by acquiring costly amino acids and vitamins from the environment.

Conclusions

Whereas successful soil microbes are predicted to have large genomes (Konstantinidis and Tiedje, 2004; Barberán et al., 2014) (Figure 3.3), *Ca.* U. copiosus has a small genome, indicating that, similar to some aquatic microbes, minimization of cellular architecture can also represent a successful strategy for soil microbes. We do not know if other uncultivated abundant soil taxa also contain reduced genomes because pre-existing genome databases are preferentially biased towards cultivated isolates. For example, only 4.5% of bacterial genomes in IMG are from uncultivated taxa (accessed April 2016). Bacteria encoding for greater metabolic versatility likely have larger genomes and therefore may be easier to cultivate in the laboratory (Button and Robertson, 2001). On the other hand, specific and combinatorial nutrient requirements such as those described for *Ca.* U. copiosus present a complex problem for researchers attempting to cultivate microbes with reduced genomes (Carini et al., 2013). Although *Ca.* U. copiosus has not yet been grown in the laboratory, cultivation is clearly a crucial next step to describing this organism, using the information described here to 'tailor' a growth medium specifically for *Ca.* U. copiosus and related microbes. Such an approach could improve our ability to describe and study the majority of soil microbes, even dominant soil microbes like *Ca.* U. copiosus, which remain difficult to cultivate under laboratory conditions.

CHAPTER IV
UNLINKED RRNA GENES ARE WIDESPREAD AMONG ENVIRONMENTAL BACTERIA AND ARCHAEA

Abstract

Ribosomes are essential to cellular life. When a complex is essential to the function of a cell, it is usually highly resistant to change and evolutionarily stable. For example, the RNA components of bacterial and archaeal ribosomes are typically organized into a single operon. This arrangement allows each rRNA component to be regulated, transcribed, and processed together - a feature thought to be important for fast and efficient growth. In reality, there are some prokaryotes that do not share this canonical rRNA order - their 16S and 23S rRNA genes are not co-located, but are instead separated and referred to as "unlinked". Such unlinked rRNA genes have previously been treated as rare exceptions or byproducts of genome degradation in intracellular bacteria. However, using a dataset of over 10,000 complete genomes, we show that unlinked rRNA genes are present in many free-living, environmental prokaryotes - most significantly within the phyla Deinococcus-Thermus, Chloroflexi, Planctomycetes, and Euryarchaeota. Using shotgun metagenomic data generated using long-read sequencing technologies, we also show that unlinked rRNA genes are common among uncultured, environmental prokaryotic populations, with up to 41% of taxa, even dominant taxa, found in soil having unlinked rRNA genes. Those environments, like soil, that presumably have slower-growing taxa tend to have far higher percentages of taxa with unlinked rRNA genes compared to environments like the human gut, where faster growing taxa are expected to predominate and unlinked rRNA genes were rarely detected. Together these results suggest that bacteria and archaea with unlinked rRNA genes are widespread and not merely atypical anecdotes. Rather, unlinked rRNA genes may confer selective advantages in some environments, but the specific nature of these advantages remains undetermined and worthy of further investigation.

Introduction

Ribosomes are the archetypal “essential proteins”, so much so that they are a key criteria in the division between cellular and viral life (Raoult and Forterre, 2008). In bacteria and archaea, the rRNA genes encoding the RNA components of the ribosome are traditionally arranged in a single operon in the order 16S - 23S - 5S. The rRNA operon is transcribed into a single RNA precursor called the pre-rRNA 30S, which is separated and processed by a number of RNases (Srivastava and Schlessinger, 1990). This arrangement of rRNA genes within a single operon is thought to be important in allowing rapid responses to changing growth conditions - the production of rRNA under a single promoter allows consistent regulation and conservation of stoichiometry between all three, essential components (Condon et al., 1995). Indeed, the production of rRNA is the rate-limiting step of ribosome synthesis (Gourse et al., 1996), and fast-growing prokaryotes can accelerate ribosome synthesis by encoding multiple rRNA operons (Klappenbach et al., 2000).

Although perhaps counter-intuitive, some prokaryotes have “unlinked” rRNA genes, with the 16S and 23S separated by large swaths of genomic space. This unlinked rRNA gene arrangement was first discovered in the thermophilic bacterium Thermus thermophilus (Hartmann et al., 1987). Reports of unlinked rRNA genes soon followed in additional bacteria, including the planctomycete Pirellula marina (Liesack and Stackebrandt, 1989), the aphid endosymbiont Buchnera aphidicola (Munson et al., 1993), and the intracellular pathogen Rickettsia prowazekii (Andersson et al., 1995). Though unlinked rRNA genes were first discovered in a free-living environmental bacterium, their ubiquity among the order Rickettsiales has led to an association between unlinked rRNA genes and the genome degradation typical of obligate intracellular lifestyles (Rurangirwa et al., 2002; Merhej et al., 2009; Andersson and Andersson, 1999).

With this study we sought to determine how common unlinked rRNA genes are across bacteria and archaea - is this unique genomic feature largely confined to those prokaryotes with an obligate intracellular lifestyle, or is it also commonly observed among environmental, free-living prokaryotes? We examined the rRNA operons of over 10,000 publicly available complete bacterial and archaeal genomes to identify which taxa have unlinked rRNA genes and to determine if there are any genomic characteristics shared across genomes with unlinked rRNA genes. As complete genomes are not typically available for the broader diversity of prokaryotes found in environmental samples (Zhi et al., 2012), we also characterized rRNA gene arrangements using long-read metagenomic datasets (Nanopore and Illumina Synthetic Long-Read Sequencing i.e. Moleculo) obtained from a range of environmental samples, which together encompassed over 17 million sequences. With these long-read metagenomic datasets, we were able to determine whether unlinked rRNA genes are common in environmental populations and how the distributions of unlinked rRNA genes differ across prokaryotic lineages and across distinct microbial habitats.

Results

Complete genome dataset

We searched all complete bacterial and archaeal genomes available on NCBI to determine how frequently "unlinked" 16S and 23S rRNA genes occur. From our set of 12240 "complete" bacterial and archaeal genomes available on NCBI as of Jan 2019, we calculated the distance between the end of the 16S rRNA gene and the beginning of the 23S rRNA gene for each rRNA gene pair. For this classification scheme, we called rRNA genes "unlinked" if this distance was greater than or equal to 1500 bp. We chose 1500 bp as our cutoff because the distance between genes in an operon is usually quite low (peaking between -20 and 30 bp in most genomes; Moreno-Hagelsieb and Collado-Vides, 2002). While we found it was common for other genes to be located between the 16S and 23S rRNA, these were most often tRNA genes, which are usually quite small, ranging from 75 to 90 bp in length (Shepherd and Ibba, 2015). When we created a histogram of the distance between each 16S and 23S rRNA gene pair from our complete genomes, we found that the vast majority of 16S and 23S rRNA pairs were < 1500bp apart (57833 out of 59496 rRNA gene pairs, Figure 4.2A).

After classifying each rRNA gene pair as linked or unlinked based on the distance between the 16S and 23S rRNA genes, we found that 4.86% of the genomes in our dataset had exclusively unlinked rRNA genes, 1.04% had mixed operons (i.e. genomes with multiple rRNA copies that had at least one unlinked rRNA gene and at least one linked operon), and 94.1% had exclusively linked operons (while 57833 out of 59496 rRNA gene pairs are linked, each genome has a variable rRNA copy number, meaning these numbers are not exactly comparable). Unlinked rRNA genes were not distributed randomly across these genomes; genomes with unlinked rRNA genes were typically from closely related lineages. We found unlinked genomes to be common (present in ≥50% of members) in taxa characterized as having an obligate intracellular lifestyle within the phyla Spirochaetes (order Spirochaetales), Epsilonproteobacteria (family Helicobacteraceae), Alphaproteobacteria (order Rickettsiales), and Tenericutes (species Mycoplasma gallisepticum). However, we also found high proportions of unlinked rRNA genes in phyla that are generally considered to be free-living, such as Deinococcus-Thermus (families Thermaceae and Deinococcaceae), Chloroflexi (family Dehalococcoidaceae), Planctomycetes (families Phycisphaeraceae and Planctomycetaceae), and Euryarchaeota (class Thermoplasmata). Phyla with at least 5% of genomes featuring exclusively unlinked rRNA genes are shown in Figure 4.1A.

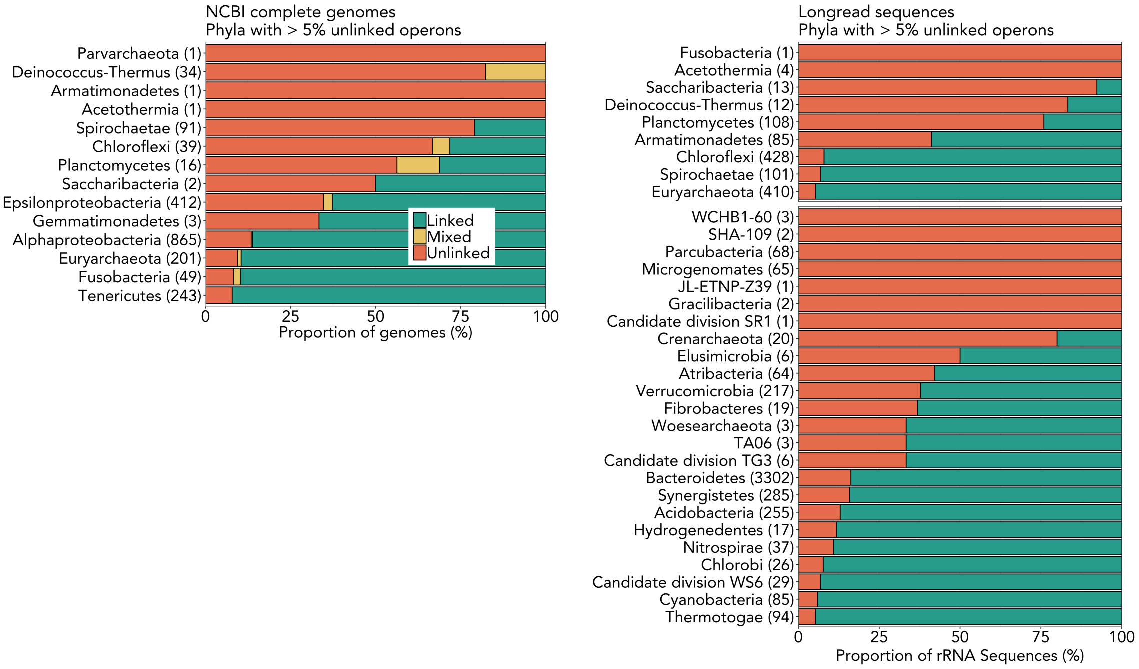


Figure 4.1: Unlinked rRNA genes occur regularly in over 30 phyla. **A**) Within a set of complete genomes from NCBI, 12 phyla had genome containing at least one unlinked rRNA operon in >5% of members. Linked refers to genomes with exclusively linked rRNA genes, unlinked refers to genomes with exclusively unlinked rRNA genes, and mixed refers to genomes with at least one of each linked and unlinked rRNA genes. **B**) Within a set of long-read sequences, we confirmed 7 of the phyla in complete genomes, and added an additional 26 phyla in which >5% of sequences were unlinked.

Long-read shotgun metagenomic dataset

While the results from our complete genome dataset demonstrated that unlinked rRNA genes are common in some free-living phyla, databases featuring complete genomes do not capture the full breadth of microbial diversity and are heavily biased towards organisms relevant to human health (Zhi et al., 2012). Just three phyla (Proteobacteria, Firmicutes, Actinobacteria) account for >85% of the genomes in our dataset - even though current estimates of bacterial diversity total 99 unique phyla (Parks et al., 2018). To investigate the ubiquity of unlinked rRNA genes among those taxa underrepresented in ‘complete’ genome databases, we analyzed long-read shotgun metagenomic data from a range of distinct sample types. Focusing on exclusively long read sequences allowed us to cover the 1500 bp distance required for classification of rRNA genes as linked or unlinked without the need for assembly. The repetitive structure of rRNA genes make them difficult to accurately assemble from the short reads typical of most current metagenomic sequencing efforts (Yuan et al., 2015).

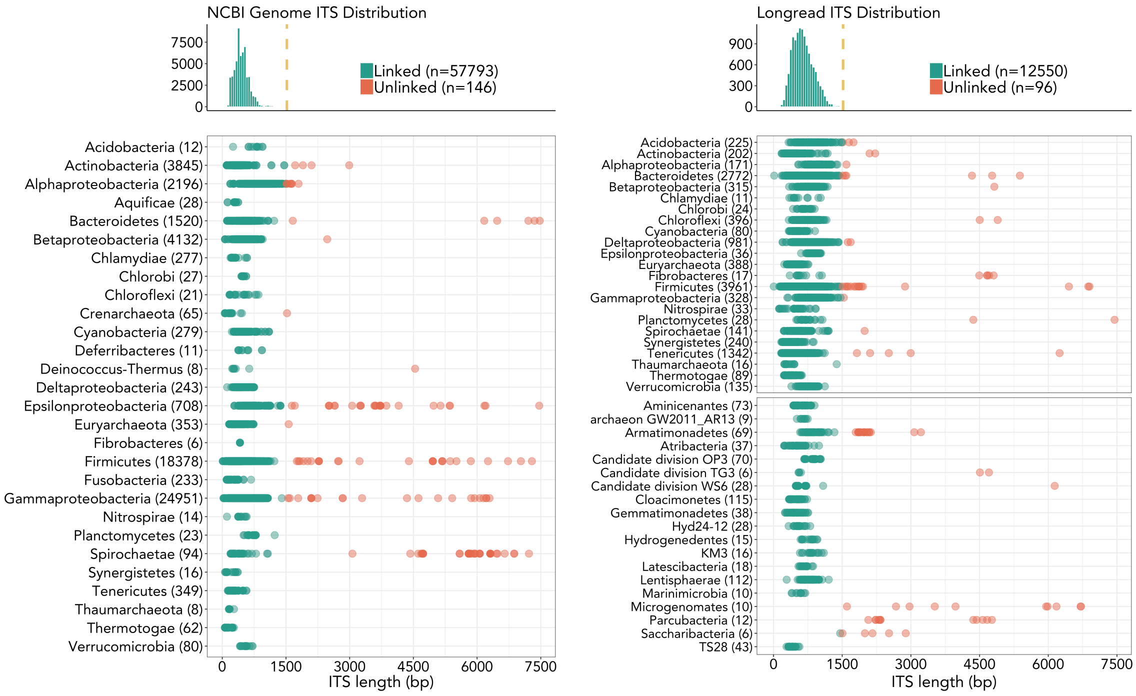


Figure 4.2: The majority of rRNA genes are linked and most have an internally transcribed spacer (ITS) < 1500bp. **A**) Distribution of ITS in complete genomes from NCBI. 97.2% of rRNA operons have an ITS < 1500bp (57833/59496). 16S and 23S separated by more than 7500 bp are not shown (1516/59496 = 2.5%). **B**) Distribution of ITS in long-read sequence dataset. 90.5% of rRNA operons have an ITS < 1500bp (12618/13932). Once again, sequences which included both a 16S and 23S rRNA but had an ITS < 7500bp are not shown (1223/13932 = 8.8%).

Out of our initial long-read dataset (~890 thousand Illumina synthetic long reads and ~16 million Nanopore sequences, with median read lengths of 7485 and 5075, respectively), only 13,932 sequences contained rRNA genes and met the criteria we established for the classification of rRNA genes as linked or unlinked (see methods). Of these sequences, we classified 1314 as unlinked, or 9.4% of the dataset. Many of the sequences classified as unlinked belonged to the same phyla where unlinked rRNA genes were prevalent in the complete genome dataset (Figure 4.1). The long-read metagenomic dataset confirmed that members of the phyla Deinococcus-Thermus, Planctomycetes, Chloroflexi, Spirochaetes, and Euryarchaeota frequently have unlinked rRNA genes (Figure 4.1B). The long-read dataset allowed us to provide additional evidence for unlinked rRNA genes in poorly studied phyla that were represented by only a handful of genomes in our complete genomes dataset, such as Acetothermia (1 genome and 4 long-read sequences) and Saccharibacteria (2 genomes and 13 long-read sequences).

The metagenomic analyses also allowed us to identify 26 additional phyla where unlinked rRNA genes are prevalent, including several candidate phyla (WCHB1-60, SHA-109, JL-ETNP-Z39, SR1, TA06, TG3, and WS6) and members of the CPR (Microgenomates and Parcubacteria, Figure 4.1). We found several clades with exclusively unlinked rRNA genes that had no representation in our complete genomes, including all Parcubacteria (68/68) and Microgenomates (65/65), Verrucomicrobia DA101 soil group (79/79), Bacteroidetes family GZKB124 (14/14), Acidobacteria Subgroup 2 (26/26), Planctomycetes order MSBL9 (23/23) and WD2101 soil group (9/9), and Chloroflexi class GIF9 (9/9). Interestingly, our long-read metagenomic analyses show that unlinked rRNA genes do not seem to be equally distributed across all environments. Some environments had higher proportions of unlinked rRNA genes - listed in descending order: soil (13-41%), sediment (7.7-29%), anaerobic digesters (8.1-8.8%) and human gut (0%), (Figure 4.3). A large portion of the unlinked rRNA genes in our soil samples belonged to the Verrucomicrobia DA101 soil group, which is one of the more abundant and widespread groups of bacteria found in soil (Brewer et al., 2016).

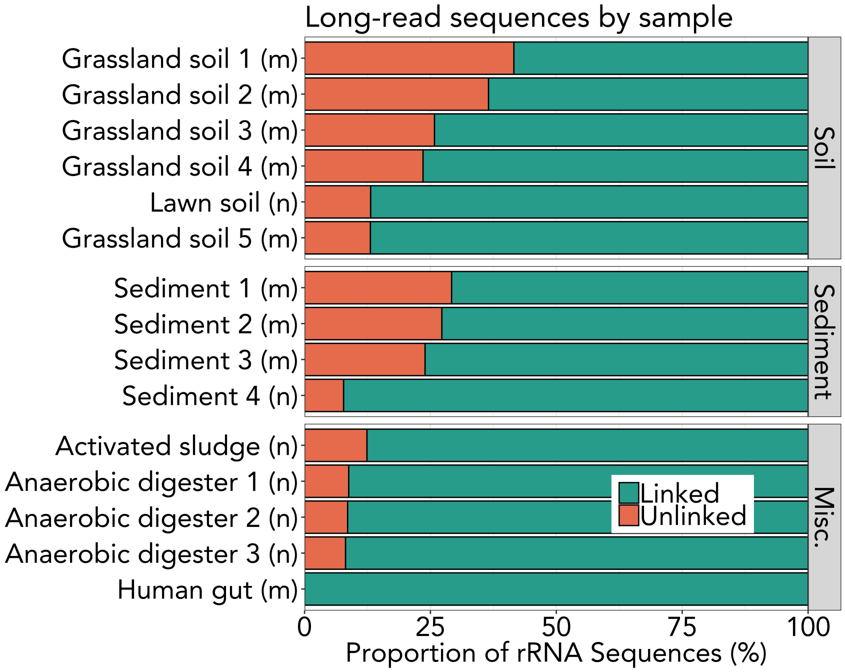


Figure 4.3: Some environments have high proportions of unlinked rRNA genes. We found soils (13-41% unlinked) and sediments (7.7-29%) to have more unlinked rRNA genes on average than anaerobic digesters (8.1-8.8%) and the human gut (0%). Moleculo sequences are indicated with an (m); nanopore sequences with an (n).

Genomic attributes associated with unlinked rRNA genes

Given that there are numerous bacterial and archaeal lineages where unlinked rRNA genes are commonly observed, we next sought to determine what other genomic features may be associated with this non-standard rRNA gene arrangement. On average, genomes with exclusively unlinked rRNA genes had fewer rRNA copies (Supplemental Figure S4.1A, χ^2 p<0.001, means of groups: 4.2 linked, 5.5 mixed, 2.6 unlinked). While genomes with unlinked rRNA genes also had smaller genomes on average, this difference was not significant (Supplemental Figure S4.1B, χ^2 p = 0.87, means of groups: 4.1Mbp linked, 4.0 Mbp mixed, 2.8 Mbp unlinked). We also calculated ΔENC' for each complete genome - a measure of codon usage bias that is negatively correlated with minimum generation time in bacteria and archaea (Vieira-Silva and Rocha, 2009). Interestingly, genomes with exclusively unlinked rRNA genes were predicted to have a longer generation times and slower potential growth (Supplemental Figure S4.1C, χ^2 p<0.001, means of groups: 0.23 linked, 0.23 mixed, 0.19 unlinked).

We also checked to see if genomes with unlinked rRNA genes were more likely to have rRNA genes that are divergent in sequence. This question is complicated by the fact that intragenomic rRNA sequence divergence becomes more common as rRNA copy number increases (Větrovský and Baldrian, 2013). We confirmed this in our dataset - the proportion of non-identical rRNA sequences within each genome was strongly correlated with rRNA copy number in both the 16S ($p < 0.001$, rho = 0.87, Pearson) and 23S ($p < 0.001$, rho = 0.90, Pearson). Additionally, in our dataset, genomes with exclusively unlinked rRNA genes tended to have lower overall rRNA copy numbers (84% had ≤3 rRNA copies). Therefore, we compared the sequence identity of 16S and 23S rRNA within every genome with 2 or 3 rRNA copies. We found that in all cases, rRNA within unlinked genomes were more dissimilar than rRNA within linked genomes (χ^2 $p < 0.05$ for all cases). However, the magnitude of these differences was not huge - in the most drastic case, in genomes with 2 rRNA copies unlinked 16S rRNA genes had an average of 5.1 mismatches versus an average of 0.71 mismatches in linked 16S rRNA genes (Supplemental Figure S4.3).

Finally, we checked if there was any connection between unlinked rRNA genes and the presence of RNaseIII genes. RNaseIII is responsible for the initial separation of the 16S and 23S rRNA transcripts once they have been transcribed into the 30S pre-rRNA (Srivastava and Schlessinger, 1990). RNaseIII is not an essential protein in most prokaryotes and several phyla in which unlinked rRNA genes are common do not encode RNaseIII (e.g. Deinococcus-Thermus & Euryarchaeota; Durand et al., 2012). Interestingly, we found that taxa with unlinked rRNA genes were significantly less likely to encode the bacterial form of RNaseIII genes (Supplemental Figure S4.3, PF00636: χ^2 $p < 0.001$, means of groups: 1.0 linked, 0.84 mixed, 0.74 unlinked; PF14622: χ^2 $p < 0.001$, means of groups: 0.86 linked, 0.63 mixed, 0.65 unlinked). We also checked this relationship for archaeal RNaseIII, but found no significant association (Supplemental Figure S4.3, PF11469: χ^2 $p = 0.153$).

Discussion

While unlinked rRNA genes have been documented previously, we have demonstrated that they are more widespread among prokaryotes than previously reported. We found that unlinked rRNA genes, whereby the 16S and 23S rRNA genes are not in close proximity within the canonical operon arrangement, consistently occur in 12 phyla using a dataset of complete genomes (Figure 4.1A), and 26 additional phyla using a dataset of metagenomic long-read sequences from five disparate environments (Figure 4.1B). Some phyla were classified as exclusively linked in our complete genome dataset, yet had many members with unlinked rRNA genes in our long-read dataset. For example, there were no complete genomes in the phylum Verrucomicrobia with unlinked rRNA genes (0/32), but in our long-read data 38% of Verrucomicrobial rRNA sequences were unlinked (82/217, most closely related to the bacterium *Ca.* Udaeobacter copiosus from the DA101 soil group; Brewer et al., 2016).

We found that taxa with unlinked rRNA genes are not randomly distributed across prokaryotic lineages - rather, we observed a strong phylogenetic signal in rRNA operon structure. To drive this point home, we assembled a phylogenetic tree from full-length 16S rRNA gene sequences from both the complete genome dataset and the long-read metagenomic dataset, where we found clusters of related taxa with exclusively unlinked rRNA genes (Figure 4.4). These lineages include: Euryarchaeota class Thermoplasmata, miscellaneous Crenarchaeota group, the vast majority of Deinococcus-Thermus, CPR divisions Parcubacteria and Microgenomates, Verrucomicrobia DA101 group, Acidobacteria subgroup 2, Chloroflexi class Dehalococcoidia, and Alphaproteobacteria class Rickettsiales. While members of the Rickettsiales are predominately obligate intracellular pathogens (Andersson and Andersson, 1999) and the CPR phyla Parcubacteria and Microgenomates contain signatures of a symbiotic lifestyle (Nelson and Stegen, 2015; Burstein et al., 2016), the rest of these clades are thought to be predominately free-living taxa.

We used our metagenomic long-read dataset to not only bypass the cultivation bias of our complete genomic dataset, but to also get an idea of the abundance of unlinked rRNA genes in environmental populations. Our analyses of long-read shotgun metagenomic datasets show that taxa with unlinked rRNA genes are far more abundant in some environments than others. Most notably, unlinked rRNA genes were much more common in soil (at the high end, 41% of rRNA genes were unlinked) than the human gut (no unlinked rRNA genes were detected). The environments with higher proportions of unlinked rRNA genes (soil & sediment) are generally thought to be populated by slow growing taxa (Brown et al., 2016; Vieira-Silva and Rocha, 2009). Likewise, we found that genomes with exclusively unlinked rRNA genes have significantly rRNA copies than genomes with mixed or exclusively linked rRNA genes, a trait which is inversely correlated with potential growth rate (Vieira-Silva and Rocha, 2009). We also found that genomes with exclusively unlinked rRNA genes are predicted to have significantly longer generation times (via the growth rate proxy ENC') compared to genomes with linked or mixed rRNA genes. These lines of evidence suggest that unlinked rRNA genes are more likely to be maintained in the genomes of taxa with slower growth rates, an observation that would need to be directly tested.

One obvious ramification of the prevalence of unlinked rRNA genes in environmental samples relates to bacterial genotyping using the full or near-full rRNA operon. While including the ITS region of the rRNA operon can increase taxonomic resolution and allow strain level identification (Zeng et al., 2012), our work shows that amplicon studies dependent on 16S and 23S rRNA genes located in close proximity may miss a large portion of bacterial and archaeal diversity. The median distance between unlinked 16S and 23S rRNA genes in our complete genome dataset was ~30kb, an impractical distance to amplify in a high-throughput manner. While strategies which use reads spanning the 16S and 23S rRNA genes to improve taxonomic resolution (e.g. Zeng et al., 2012; Cuscó et al., 2018) are less likely to be biased in some environments (e.g. human gut), they will likely miss many phylogenetic groups in other environments like soil and sediment, with a potential loss of up to 41% of rRNA genes.

Upon first consideration, unlinking the 16S and 23S rRNA genes would seem to be disadvantageous given that both rRNA are needed in equal proportions in the final ribosome. While we do not know how unlinked rRNA genes might affect cell fitness, it seems unlikely this non-canonical rRNA gene arrangement has a substantial negative effect in the environmental, free-living taxa in which it occurs. Free-living organisms are generally under greater selective pressure than obligate intracellular organisms and are less prone to the fixation of deleterious mutations; species with large effective population sizes (N_e) face strong selection and weak genetic drift (Batut et al., 2014). Therefore, it seems that if this gene rearrangement, which affects arguably the most important complex for cellular life, had a substantial negative effect it would not persist across so many free-living groups, or in specific lineages that can be incredibly abundant in some environments (e.g. the Verrucomicrobia *Ca.* U. copiosus; Brewer et al., 2016). Additionally, taxa with unlinked rRNA genes are likely under less pressure to maintain optimal efficiency in the production of rRNA, as we have shown they seem to have slower growth rates than taxa with traditional rRNA operons. Moreover, studies in E.coli have shown that unbalanced rRNA gene dosage does not lead to severe consequences- balanced synthesis of ribosomal proteins still occurs and excess rRNA is rapidly degraded (Siehnel and Morgan, 1985). While the doubling time of E. coli with unbalanced rRNA stoichiometry did increase in this study (by 40 minutes on average), few environmental bacteria ever achieve the growth sprints E. coli is capable of - turnover times of saprotrophic soil bacterial communities can be on the order of weeks (Rousk and Bååth, 2011).

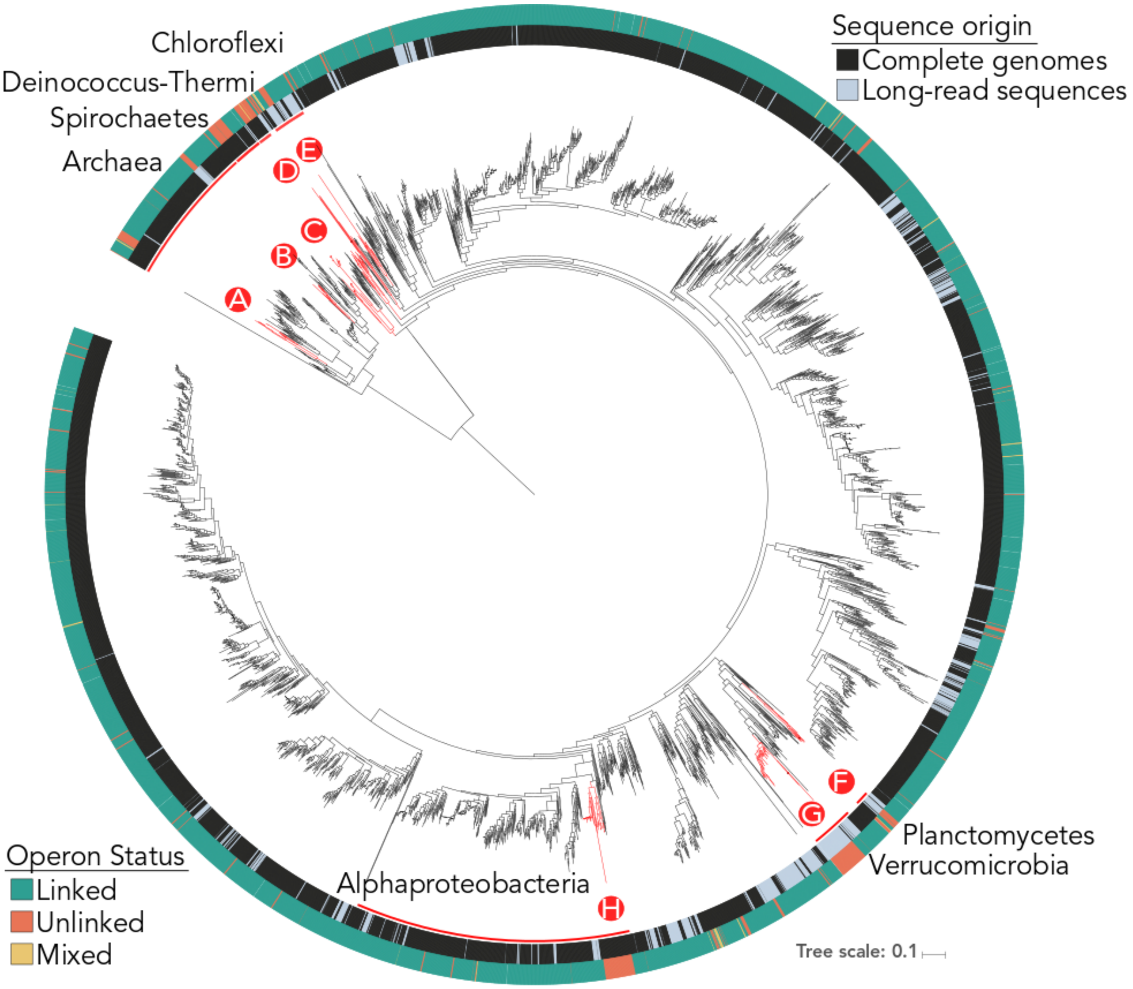


Figure 4.4: Unlinked rRNA genes are phylogenetically conserved. A phylogenetic tree created from full-length 16S rRNA sequences by combining both the NCBI complete genome and long-read datasets. We included only one 16S rRNA gene from each unique species within the complete genomes dataset. To represent our long-read metagenomic dataset, we matched each partial 16S rRNA extracted by metaxa2 to full-length sequences in the SILVA 132 SSU database. We added full-length SILVA 16S rRNA sequences as representatives of long-read sequences if they matched to at least 95% identity. The inner ring indicates which dataset each sequence originated from, while the outer ring indicates the status of rRNA genes as either linked, mixed, or unlinked. Sequence representatives of the long-read dataset cannot be mixed,

as there was no way for us to distinguish multi-copy rRNA genes. Phyla with significant proportions of unlinked members are indicated with text while clades with exclusively unlinked members are colored red: A) Euryarchaeota class Thermoplasmata, B) Miscellaneous Crenarchaeotic Group, C) Spirochaetae class Spirochaetes, D) Deinococcus-Thermi, E) Chlorflexi class Dehalococcoidia, F) Planctomycetes families Phycisphaeraceae and Planctomycetaceae, G) Verrucomicrobia DA101 soil group and H) Alphaproteobacteria order Rickettsiales.

While we do not know if unlinked rRNA genes have an effect on the fitness of their owners, we can speculate on hypothetical benefits this rearrangement might confer. By transcribing the 16S and 23S rRNA genes separately, taxa with unlinked rRNA genes may eliminate or reduce the need for RNaseIII, the ribonuclease that separates these segments after they are transcribed into the classical 30S pre-rRNA. Indeed, we found less evidence of bacterial RNaseIII in taxa with unlinked rRNA genes (Supplemental Figure S4.3), including a complete absence of the protein in the phyla Deinococcus-Thermi and Gemmatimonadetes. Interestingly, some bacteriophages leverage host RNase III to process their own mRNA (Gone et al., 2016); in some species, the presence of RNase III can stimulate the translation of infecting phage mRNA by several orders of magnitude (Wilcon et al., 2002), although other phage appear indifferent to the presence of RNase III (Hagen and Young, 1978). Regardless, it seems that the loss of RNaseIII is not without negative consequences in organisms with unlinked rRNA genes - recent work has shown that knocking out RNaseIII in *Borrelia burgdorferi* (a spirochete with unlinked rRNA) did not affect the processing of the 16S or 5S rRNA, but did affect the maturation of the 23S and resulted in a decreased growth rate and increased cell length (Anacker et al., 2018).

Interestingly, we found that in genomes with 2-3 rRNA copies, genomes with unlinked rRNA genes had greater intragenomic rRNA divergence than genomes with exclusively linked rRNA genes. In other words, genomes with unlinked rRNA genes had greater sequence divergence between their rRNA gene copies. While differential expression of divergent rRNA has been observed in *Streptomycetes coelicolor* (Kim et al., 2007) and *E. coli* (Condon et al., 1992), it was previously unclear whether this phenomenon had any effect on the translational activity of the ribosome (Byrgazov et al., 2013). However, recent work has shown that divergent rRNAs can regulate gene expression in *Vibrio vulnificus*, with ribosomes composed of the most divergent rRNAs preferentially translating a set of mRNAs related to temperature and nutrient shifts (Song et al., 2019). In this study, the ribosomes with altered activity contained only 3 divergent nucleotides in the 16S and 16 in the 23S rRNA. Splitting the rRNA operon could allow differential expression between 16S and 23S rRNA genes that were previously co-transcribed, leading to further customization of heterogeneous ribosomes with potentially altered activity.

Conclusions

While we do not know why unlinked rRNA genes are so prevalent (particularly in these bacteria and archaea found in environmental samples for which complete genomes are not yet available), we have shown that this rearrangement appears to occur more frequently in taxa with slower predicted growth rates and may be related to the presence of RNase III or divergent rRNA. Regardless, we have shown that up to 41% of rRNA genes in some environments are unlinked - meaning unlinked rRNA genes are far from atypical anecdotes. We have developed a number of hypotheses about potential advantages of unlinked rRNA operons that could be tested experimentally - especially as a number of taxa with unlinked rRNA operons are relatively easy to manipulate in culture (Holland et al., 2006; Devos, 2013).

CHAPTER V
CONCLUSIONS

My thesis research has shed a measure of light on the uncultivated majority dwelling in soil - a poorly understood population responsible for key ecosystem processes in soil that harbors a rich and untapped genetic diversity. In Chapter II I showed that as depth increases in a soil profile, the proportion of uncultivated, novel bacteria and archaea increases. In nutrient poor deep soils, I showed that candidate phyla increase in abundance and assembled two genomes from metagenomic data from one such phylum - candidate phylum AD3.

In Chapter III I described one bacterium in particular, *Candidatus* Udaeobacter copiosus, which dominates soil communities around the globe. I analyzed a representative genome of this species, and found that this bacterium has a significantly smaller genome than the typical soil dweller, and that this small genome is rife with putative auxotrophies. Although *Ca.* U. copiosus must depend on its environment for various amino acids and vitamins, it is highly successful in terms of ubiquity and relative abundance, demonstrating that taxa with reduced genomes can be successful in soil. In Chapter IV I describe a peculiar rRNA gene rearrangement that was previously linked to obligate intracellular bacteria and thought to be a consequence of genome degradation. I show that this rearrangement occurs in a number of environmental, free-living taxa and is widespread in many environmental samples - I found that up to 41% of rRNA genes in one soil were unlinked. This rearrangement also occurs in evolutionarily "successful" bacteria - notably *Ca.* U. copiosus. This research has shown that while most of our understanding of bacterial cells is derived from model organisms like *Bacillus subtilis* and *Escherichia coli.*, these species are not representative of the majority of bacterial life on this planet, and that peering into the uncultured majority in environmental samples can reveal interesting life history strategies and unique biology.

REFERENCES

Akashi H, Gojobori T. (2002). Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proceedings of the National Academy of Sciences* 99: 1–6.

Anacker ML, Drecktrah D, LeCoultre RD, Lybecker M, Samuels DS. (2018). RNase III Processing of rRNA in the Lyme Disease Spirochete *Borrelia burgdorferi*. Henkin TM (ed). *Journal of Bacteriology* 200: 1–11.

Anantharaman K, Breier JA, Dick GJ. (2016). Metagenomic resolution of microbial functions in deep-sea hydrothermal plumes across the Eastern Lau Spreading Center. *ISME Journal* 10: 225–239.

Andersson JO, Andersson SGE. (1999). Genome Degradation is an Ongoing Process in *Rickettsia*. *Molecular Biology and Evolution* 16: 1178–1191.

Andersson SGE, Zomorodipour A, Winkler HH, Kurland CG. (1995). Unusual Organization of the rRNA Genes in *Rickettsia prowazekii*. *Journal of Bacteriology* 177: 4171–4175.

Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* 3: e1029–17.

Baker BJ, Saw JH, Lind AE, Lazar CS, Hinrichs K-U, Teske AP, *et al.* (2016). Genomic inference of the metabolism of cosmopolitan subsurface Archaea, Hadesarchaea. *Nature Microbiology* 1: 1–7.

Balesdent J, Basile-Doelsch I, Chadoeuf J, Cornu S, Derrien D, Fekiacova Z, *et al.* (2018). Atmosphere–soil carbon transfer as a function of soil depth. *Nature* 559: 599–602.

Banning NC, Maccarone LD, Fisk LM, Murphy DV. (2015). Ammonia-oxidising bacteria not archaea dominate nitrification activity in semi-arid agricultural soil. *Nature* 5: 1–8.

Barberán A, Ramirez KS, Leff JW, Bradford MA, Wall DH, Fierer N. (2014). Why are some microbes more ubiquitous than others? Predicting the habitat breadth of soil bacteria Klironomos J (ed). *Ecology Letters* 17: 794–802.

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

Bengtsson-Palme J, Hartmann M, Eriksson KM, Pal C, Thorell K, Larsson DGJ, *et al.* (2015). metaxa2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Mol Ecol Resour* 15: 1403–1414.

Bergmann GT, Bates ST, Eilers KG, Lauber CL, Caporaso JG, Walters WA, *et al.* (2011). The under-recognized dominance of Verrucomicrobia in soil bacterial communities. *Soil Biology and Biochemistry* 43: 1450–1455.

Billings SA, Hirmas D, Sullivan PL, Lehmeier CA, Bagchi S, Min K, *et al.* (2018). Loss of deep roots limits biogenic agents of soil development that are only partially restored by decades of forest regeneration. *Elementa* 6: 1–19.

Bissett A, Fitzgerald A, Meintjes T, Mele PM, Reith F, Dennis PG, *et al.* (2016). Introducing BASE: the Biomes of Australian Soil Environments soil microbial diversity database. *GigaScience* 5: 1–11.

Blume E, Bischoff M, Reichert JM, Moorman T, Konopka A, Turco RF. (2002). Surface and subsurface microbial biomass, community structure and metabolic activity as a function of soil depth and season. *Applied Soil Ecology* 20: 171–181.

Brewer TE, Handley KM, Carini P, Gilbert JA, Fierer N. (2016). Genome reduction in an abundant and ubiquitous soil bacterium '*Candidatus* Udaeobacter copiosus'. *Nature Microbiology* 2: 16198.

Brown CT, Olm MR, Thomas BC, Banfield JF. (2016). Measurement of bacterial replication rates in microbial communities. *Nat Biotechnol* 34: 1256–1263.

Browne HP, Forster SC, Anonye BO, Kumar N, Neville BA, Stares MD, *et al.* (2016). Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature* 533: 543–546.

Burstein D, Sun CL, Brown CT, Sharon I, Anantharaman K, Probst AJ, *et al.* (2016). Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nature Communications* 7: 1–8.

Button DK, Robertson BR. (2001). Determination of DNA content of aquatic bacteria by flow cytometry. *Applied and Environmental Microbiology* 67: 1636–1645.

Byrgazov K, Vesper O, Moll I. (2013). Ribosome heterogeneity: another level of complexity in bacterial translation regulation. *Current Opinion in Microbiology* 16: 133–139.

Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. (2010a). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* 26: 266–267.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, *et al.* (2010b). QIIME allows analysis of high-throughput community sequencing data. *Nat Meth* 7: 335–336.

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

Carini P, Steindler L, Beszteri S, Giovannoni SJ. (2013). Nutrient requirements for growth of the extreme oligotroph '*Candidatus* Pelagibacter ubique' HTCC1062 on a defined medium. *ISME Journal* 7: 592–602.

Cavaletti L, Monciardini P, Bamonte R, Schumann P, Rohde M, Sosio M, *et al.* (2006). New lineage of filamentous, spore-forming, gram-positive bacteria from soil. *Applied and Environmental Microbiology* 72: 4360–4369.

Chen I-MA, Markowitz VM, Chu K, Palaniappan K, Szeto E, Pillay M, *et al.* (2017). IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Research* 45: D507–D516.

Ciccarelli FD, Doerks T, Mering von C, Creevey CJ, Snel B, Bork P. (2006). Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *Science* 311: 1283–1287.

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, *et al.* (2013). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research* 42: D633–D642.

Condon C, Philips J, Fu Z-Y, Squires C, Squires CL. (1992). Comparison of the expression of the seven ribosomal RNA operons in Escherichia coli. *The EMBO Journal* 11: 4175–4185.

Condon C, Squires C, Squires CL. (1995). Control of rRNA Transcription in *Escherichia coli*. *Microbiological Reviews* 59: 623–645.

Crits-Christoph A, Diamond S, Butterfield CN, Thomas BC, Banfield JF. (2018). Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. *Nature* 1–17.

Crowther TW, Maynard DS, Leff JW, Oldfield EE, McCulley RL, Fierer N, *et al.* (2014). Predicting the responsiveness of soil biodiversity to deforestation: a cross-biome study. *Glob Change Biol* 20: 2983–2994.

Cuscó A, Catozzi C, Viñes J, Sanchez A, Francino O. (2018). Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and whole rrn operon. *F1000Res* 7: 1755–25.

D'Souza G, Waschina S, Pande S, Bohl K, Kaleta C, Kost C. (2014). Less is more: selective advantages can explain the prevalent loss of biosynthetic genes in bacteria. *Evolution* 68: 2559–2570.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, *et al.* (2006). Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Applied and Environmental Microbiology* 72: 5069–5072.

Devos DP. (2013). Gemmata obscuriglobus. *CURBIO* 23: R705–R707.

Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, *et al.* (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biol* 10: 1–16.

Dunfield PF, Yuryev A, Senin P, Smirnova AV, Stott MB, Hou S, *et al.* (2007). Methane oxidation by an extremely acidophilic bacterium of the phylum Verrucomicrobia. *Nature* 450: 879–882.

Durand S, Gilet L, Condon C. (2012). The Essential Function of B. subtilis RNase III Is to Silence Foreign Toxin Genes Viollier PH (ed). *PLOS Genetics* 8: e1003181–11.

Eddy SR. (2011). Accelerated Profile HMM Searches Pearson WR (ed). *PLoS Comput Biol* 7: e1002195–16.

Edgar RC. (2010a). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460–2461.

Edgar RC. (2010b). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460–2461.

Edgar RC. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv* 1–21.

Edgar RC. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Meth* 10: 996–998.

Eilers KG, Debenport S, Anderson S, Fierer N. (2012). Digging deeper to find unique microbial communities: The strong effect of depth on the structure of bacterial and archaeal communities in soil. *Soil Biology and Biochemistry* 50: 58–65.

Elharar Y, Roth Z, Hermelin I, Moon A, Peretz G, Shenkerman Y, *et al.* (2014). Survival of mycobacteria depends on proteasome-mediated amino acid recycling under nutrient limitation. *The EMBO Journal* 33: 1802–1814.

Eloe-Fadrosh EA, Ivanova NN, Woyke T, Kyrpides NC. (2016). Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nature Microbiology* 1: 1–4.

Everard A, Belzer C, Geurts L, Ouwerkerk JP, Druart C, Bindels LB, *et al.* (2013). Cross-talk between *Akkermansia muciniphila* and intestinal epithelium controls diet-induced obesity. *Proceedings of the National Academy of Sciences* 110: 9066–9071.

Farrell M, Hill PW, Farrar J, DeLuca TH, Roberts P, Kielland K, *et al.* (2013). Oligopeptides Represent a Preferred Source of Organic N Uptake: A Global Phenomenon? *Ecosystems* 16: 133–145.

Felske A, Akkermans ADL. (1998). Prominent occurrence of ribosomes from an uncultured bacterium of the Verrucomicrobiales cluster in grassland soils. *Lett Appl Microbiol* 26: 219–223.

Fierer N. (2017). Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat Rev Micro* 1–12.

Fierer N, Ladau J, Clemente JC, Leff JW, Owens SM, Pollard KS, *et al.* (2013). Reconstructing the Microbial Diversity and Function of Pre-Agricultural Tallgrass Prairie Soils in the united States. *Science* 342: 621–643.

Fierer N, Lauber CL, Ramirez KS, Zaneveld J, Bradford MA, Knight R. (2011). Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *ISME Journal* 6: 1007–1017.

Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, *et al.* (2012). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences* 109: 1–6.

Fierer N, Lennon JT. (2011). The generation and maintenance of diversity in microbial communities. *American Journal of Botany* 98: 439–448.

Fierer N, Schimel JP, Holden PA. (2003). Variations in microbial community composition through two soil depth profiles. *Soil Biology and Biochemistry* 35: 167–176.

Fierer N, Strickland MS, Liptzin D, Bradford MA, Cleveland CC. (2009). Global patterns in belowground communities. *Ecology Letters* 12: 1238–1249.

Fimmen RL, Richter DD Jr., Vasudevan D, Williams MA, West LT. (2008). Rhizogenic Fe–C redox cycling: a hypothetical biogeochemical mechanism that drives crustal weathering in upland soils. *Biogeochemistry* 87: 127–141.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, *et al.* (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* 44: D279–D285.

Flynn TM, Koval JC, Greenwald SM, Owens SM, Kemner KM, Antonopoulos DA. (2017). Parallelized, Aerobic, Single Carbon-Source Enrichments from Different Natural Environments Contain Divergent Microbial Communities. *Front Microbiol* 8: 1540–14.

Friedel JK, Scheller E. (2002). Composition of hydrolysable amino acids in soil organic matter and soil microbial biomass. *Soil Biology and Biochemistry* 34: 315–325.

Fujita M, Losick R. (2005). Evidence that entry into sporulation in Bacillus subtilis is governed by a gradual increase in the level and activity of the master regulator Spo0A. *Genes Development* 19: 2236–2244.

Fung T, Kwong N, van der Zwan T, Wu M. (2013). Residual Glycogen Metabolism in *Escherichia coli* is Specific to the Limiting Macronutrient and Varies During Stationary Phase. *Journal of Experimental Microbiology and Immunology JEMI* 17: 83–87.

Garcia SL, Buck M, McMahon KD, Grossart H-P, Eiler A, Warnecke F. (2015). Auxotrophy and intrapopulation complementary in the 'interactome' of a cultivated freshwater model community. *Mol Ecol* 24: 4449–4459.

Giovannoni SJ, Thrash JC, Ben Temperton. (2014). Implications of streamlining theory for microbial ecology. *ISME Journal* 8: 1553–1565.

Giovannoni SJ, Vergin KL. (2012). Seasonality in Ocean Microbial Communities. *Science* 335: 671–676.

Goldfarb KC, Karaoz U, Hanson CA, Santee CA, Bradford MA, Treseder KK, *et al.* (2011). Differential growth responses of soil bacterial taxa to carbon substrates of varying chemical recalcitrance. *Front Microbiol* 2: 1–10.

Gone S, Alfonso-Prieto M, Paudyal S, Nicholson AW. (2016). Mechanism of Ribonuclease III Catalytic Regulation by Serine Phosphorylation. *Nature* 1–9.

Gourse RL, Gaal T, Bartlett MS, Appleman JA, Ross W. (1996). rRNA Transcription and Growth Rate–Dependent Regulation of Ribosome Synthesis in *Escherichia coli*. *Annu Rev Microbiol* 50: 645–677.

Guo J, Cole JR, Zhang Q, Brown CT, Tiedje JM. (2016). Microbial Community Analysis with Ribosomal Gene Fragments from Shotgun Metagenomes. Schloss PD (ed). *Applied and Environmental Microbiology* 82: 157–166.

Hagen FS, Young ET. (1978). Effect of RNase III on Efficiency of Translation of Bacteriophage T7 Lysozyme mRNA. *Journal of Virology* 26: 793–804.

Hall SJ, Liptzin D, Buss HL, DeAngelis K, Silver WL. (2016). Drivers and patterns of iron redox cycling from surface to bedrock in a deep tropical forest soil: a new conceptual model. *Biogeochemistry* 130: 177–190.

Hartmann RK, Ulbrich N, Erdmann VA. (1987). An unusual rRNA operon constellation: in *Thermus thermophilus* HB8 the 23S/5S rRNA operon is a separate entity from the 16S rRNA operon. *Biochimie* 1097–1104.

Hayatsu M, Tago K, Saito M. (2008). Various players in the nitrogen cycle: Diversity and functions of the microorganisms involved in nitrification and denitrification. *Soil Science and Plant Nutrition* 54: 33–45.

Herlemann DPR, Lundin D, Labrenz M, Jürgens K, Zheng Z, Aspeborg H, *et al.* (2013). Metagenomic de novo assembly of an aquatic representative of the verrucomicrobial class Spartobacteria. Azam F, Simon M (eds). *mBio* 4: e00569–12.

Hille R. (2005). Molybdenum-containing hydroxylases. *Archives of Biochemistry and Biophysics* 433: 107–116.

Holland AD, Rothfuss HM, Lidstrom ME. (2006). Development of a defined medium supporting rapid growth for Deinococcus radiodurans and analysis of metabolic capacities. *Appl Microbiol Biotechnol* 72: 1074–1082.

Hou S, Makarova KS, Saw JH, Senin P, Ly BV, Zhou Z, *et al.* (2008). Complete genome sequence of the extremely acidophilic methanotroph isolate V4, Methylacidiphilum infernorum, a representative of the bacterial phylum Verrucomicrobia. *Biol Direct* 3: 26–25.

Howe AC, Jansson JK, Malfatti SA, Tringe SG, Tiedje JM, Brown CT. (2014). Tackling soil diversity with the assembly of large, complex metagenomes. *Proceedings of the National Academy of Sciences* 111: 4904–4909.

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, *et al.* (2016). A new view of the tree of life. *Nature Microbiology* 1: 1–6.

Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11: 673–11.

Janssen PH. (2006). Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Applied and Environmental Microbiology* 72: 1719–1728.

Ji M, Greening C, Vanwonterghem I, Carere CR, Bay SK, Steen JA, *et al.* (2017). Atmospheric trace gases support primary production in Antarctic desert surface soil. *Nature* 552: 400–403.

Jobbágy EG, Jackson RB. (2000). The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecological Applications* 10: 423–436.

Jones DT, Taylor WR, Thornton JM. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Application in the Biosciences* 8: 275–282.

Jørgensen TS, Kiil AS, Hansen MA, Sørensen SJ, Hansen LH. (2015). Current strategies for mobilome research. *Front Microbiol* 5: 1–6.

Kandror O, DeLeon A, Goldberg AL. (2002). Trehalose synthesis is induced upon exposure of Escherichia coli to cold and is essential for viability at low temperatures. *Proc Natl Acad Sci USA* 99: 1–6.

Kant R, van Passel MWJ, Palva A, Lucas S, Lapidus A, Glavina del Rio T, *et al.* (2011). Genome sequence of Chthoniobacter flavus Ellin428, an aerobic heterotrophic soil bacterium. *Journal of Bacteriology* 193: 2902–2903.

Kantor RS, Wrighton KC, Handley KM, Sharon I, Hug LA, Castelle CJ, *et al.* (2013). Small Genomes and Sparse Metabolisms of Sediment-Associated Bacteria from Four Candidate Phyla Brune A, Giovannoni SJ (eds). *mBio* 4: 1–11.

Khadem AF, vanTeeseling MCF, van Niftrik L, Jetten MSM, Camp den HJMO, Pol A. (2012). Genomic and physiological analysis of carbon storage in the verrucomicrobial methanotroph '*Ca.* Methylacidiphilum fumariolicum' SolV. *Front Microbiol* 3: 1–10.

Kim H-L, Shin E-K, Kim H-M, Ryou S-M, Kim S, Cha C-J, *et al.* (2007). Heterogeneous rRNAs are differentially expressed during the morphological development of Streptomyces coelicolor. *FEMS Microbiology Letters* 275: 146–152.

Kim HM, Jung JY, Yergeau E, Hwang CY, Hinzman L, Nam S, *et al.* (2014). Bacterial community structure and soil properties of a subarctic tundra soil in Council, Alaska. *FEMS*

Microbiol Ecol 89: 465–475.

Kim W, Levy SB. (2008). Increased Fitness of Pseudomonas fluorescens Pf0-1 Leucine Auxotrophs in Soil. *Applied and Environmental Microbiology* 74: 3644–3651.

King GM, Weber CF. (2007). Distribution, diversity and ecology of aerobic CO-oxidizing bacteria. *Nat Rev Micro* 5: 107–118.

Klappenbach JA, Dunbar JM, Schmidt TM. (2000). rRNA Operon Copy Number Reflects Ecological Strategies of Bacteria. *Applied and Environmental Microbiology* 66: 1328–1333.

Konstantinidis KT, Tiedje JM. (2005). Towards a genome-based taxonomy for prokaryotes. *Journal of Bacteriology* 187: 6258–6264.

Konstantinidis KT, Tiedje JM. (2004). Trends between gene content and genome size in prokaryotic species with larger genomes. *Proceedings of the National Academy of Sciences* 101: 3160–3165.

Kotak M, Isanapong J, Goodwin L, Bruce D, Chen A, Han CS, *et al.* (2015). Complete Genome Sequence of the *Opitutaceae* Bacterium Strain TAV5, a Potential Facultative Methylotroph of the Wood-Feeding Termite *Reticulitermes flavipes*. *Genome Announc* 3: 293–2.

Kramer C, Gleixner G. (2008). Soil organic matter in soil depth profiles: Distinct carbon preferences of microbial groups during carbon transformation. *Soil Biology and Biochemistry* 40: 425–433.

Kramer S, Marhan S, Haslwimmer H, Ruess L, Kandeler E. (2013). Temporal variation in surface and subsoil abundance and function of the soil microbial community in an arable soil. *Soil Biology and Biochemistry* 61: 76–85.

Kuleshov V, Jiang C, Zhou W, Jahanbani F, Batzoglou S, Snyder M. (2016). Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nat Biotechnol* 34: 64–69.

Kuleshov V, Xie D, Chen R, Pushkarev D, Ma Z, Blauwkamp T, *et al.* (2014). Whole-genome haplotyping using long reads and statistical methods. *Nat Biotechnol* 32: 261–266.

Kumar S, Stecher G, Tamura K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution* 33: 1870–1874.

Leff JW, Jones SE, Prober SM, Barberán A, Borer ET, Firn JL, *et al.* (2015). Consistent responses of soil microbial communities to elevated nutrient inputs in grasslands across the globe. *Proc Natl Acad Sci USA* 112: 10967–10972.

Letunic I, Bork P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research* 44: W242–W245.

Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31: 1674–1676.

Liesack W, Stackebrandt E. (1989). Evidence for Unlinked rrn Operons in the Planctomycete *Pirellula marina*. *Journal of Bacteriology* 171: 5025–5030.

Lin X, Kennedy D, Fredrickson J, Bjornstad B, Konopka A. (2011). Vertical stratification of subsurface microbial community composition across geological formations at the Hanford Site. *Environmental Microbiology* 14: 414–425.

Ling LL, Schneider T, Peoples AJ, Spoering AL, Engels I, Conlon BP, *et al.* (2015). A new antibiotic kills pathogens without detectable resistance. *Nature* 1–19.

Liptzin D, Silver WL. (2009). Effects of carbon additions on iron reduction and phosphorus availability in a humid tropical forest soil. *Soil Biology and Biochemistry* 41: 1696–1702.

Lloyd KG, Steen AD, Ladau J, Yin J, Crosby L. (2018). Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes Neufeld JD (ed). *mSystems* 3: 1–12.

Lochhead AG. (1958). Soil bacteria and growth-promoting substances. *Bacteriology Reviews* 22: 145–153.

Lorite MJ, Tachil J, Sanjuán J, Meyer O, Bedmar EJ. (2000). Carbon Monoxide Dehydrogenase Activity in Bradyrhizobium japonicum. *Applied and Environmental Microbiology* 66: 1871–1876.

Marty C, Houle D, Gagnon C, Courchesne F. (2017). The relationships of soil total nitrogen concentrations, pools and C:N ratios with climate, vegetation types and nitrate deposition in temperate and boreal forests of eastern Canada. *Catena* 152: 163–172.

Mee MT, Collins JJ, Church GM, Wang HH. (2014). Syntrophic exchange in synthetic microbial communities. *Proc Natl Acad Sci USA* 111: E2149–56.

Merhej V, Royer-Carenzi M, Pontarotti P, Raoult D. (2009). Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. *Biol Direct* 4: 13–25.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, *et al.* (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9: 386–8.

Miller CS, Handley KM, Wrighton KC, Frischkorn KR, Thomas BC, Banfield JF. (2013). Short-Read Assembly of Full-Length 16S Amplicons Reveals Bacterial Diversity in Subsurface Sediments Gilbert JA (ed). *PLoS ONE* 8: e56018–11.

Moreno-Hagelsieb G, Collado-Vides J. (2002). A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics* 18: S329–S336.

Morris JJ, Lenski RE, Zinser ER. (2012). The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss. *mBio* 3: 1–7.

Munson MA, Baumann L, Baumann P. (1993). *Buchnera aphidicola* (a prokaryotic endosymbiont of aphids) contains a putative 16S rRNA operon unlinked to the 23s rRNA-encoding gene: sequence determination, and promoter and terminator analysis. *Gene* 137: 171–178.

Nayfach S, Pollard KS. (2015). Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol* 16: 59–18.

Nelson WC, Stegen JC. (2015). The reduced genomes of Parcubacteria (OD1) contain signatures of a symbiotic lifestyle. *Front Microbiol* 6: 693–14.

Novembre JA. (2002). Accounting for Background Nucleotide Composition When Measuring Codon Usage Bias. *Molecular Biology and Evolution* 19: 1390–1394.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, *et al.* (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* 44: D733–D745.

Oh N-H, Richter DD. (2005). Elemental translocation and loss from three highly weathered soil–bedrock profiles in the southeastern United States. *Geoderma* 126: 5–25.

Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, *et al.* (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* 36: 996–1004.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 25: 1043–1055.

Peng Y, Leung HCM, Yiu SM, Chin FYL. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28: 1420–1428.

Portillo MC, Leff JW, Lauber CL, Fierer N. (2013). Cell size distributions of soil bacterial and archaeal taxa. *Applied and Environmental Microbiology* 79: 7610–7617.

Price MN, Dehal PS, Arkin AP. (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution* 26: 1641–1650.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, *et al.* (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* 41: D590–6.

Raes J, Korbel JO, Lercher MJ, Mering von C, Bork P. (2007). Prediction of effective genome size in metagenomic samples. *Genome Biol* 8: 1–11.

Ramirez KS, Leff JW, Barberan A, Bates ST, Betley J, Crowther TW, *et al.* (2014). Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally. *Proceedings of the Royal Society B: Biological Sciences* 281: 20141988–20141988.

Raoult D, Forterre P. (2008). Redefining viruses: lessons from Mimivirus. *Nat Rev Micro* 6: 315–319.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, *et al.* (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499: 431–437.

Roller BRK, Stoddard SF, Schmidt TM. (2016). Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. *Nature Microbiology* 1: 1–7.

Rousk J, Bååth E. (2011). Growth of saprotrophic fungi and bacteria in soil. *FEMS Microbiol Ecol* 78: 17–30.

Rurangirwa FR, Brayton KA, McGuire TC, Knowles DP, Palmer GH. (2002). Conservation of the unique rickettsial rRNA gene arrangement in *Anaplasma*. *International Journal Of Systematic And Evolutionary Microbiology* 52: 1405–1409.

Sabath N, Ferrada E, Barve A, Wagner A. (2013). Growth Temperature and Genome Size in Bacteria Are Negatively Correlated, Suggesting Genomic Streamlining During Thermal Adaptation. *Genome Biology and Evolution* 5: 966–977.

Sangwan P, Chen X, Hugenholtz P, Janssen PH. (2004). Chthoniobacter flavus gen. nov., sp. nov., the first pure-culture representative of subdivision two, Spartobacteria classis nov., of the phylum Verrucomicrobia. *Applied and Environmental Microbiology* 70: 5875–5881.

Schütz K, Kandeler E, Nagel P, Scheu S, Ruess L. (2010). Functional microbial community response to nutrient pulses by artificial groundwater recharge practice in surface soils and subsoils. *FEMS Microbiol Ecol* 72: 445–455.

Schwarz A, Adetutu EM, Juhasz AL, Aburto-Medina A, Ball AS, Shahsavari E. (2018). Microbial Degradation of Phenanthrene in Pristine and Contaminated Sandy Soils. *Microb Ecol* 75: 888–902.

Sharon I, Kertesz M, Hug LA, Pushkarev D, Blauwkamp TA, Castelle CJ, *et al.* (2015). Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. *Genome Res* 25: 534–543.

Shepherd J, Ibba M. (2015). Bacterial transfer RNAs. *FEMS Microbiol Rev* 39: 280–300.

Shimkets LJ. (1999). Intercellular signaling during fruiting-body development of *Myxococcus xanthus*. *Annu Rev Microbiol* 53: 1–26.

Siehnel RJ, Morgan EA. (1985). Unbalanced rRNA Gene Dosage and its Effects on rRNA and Ribosomal-Protein Synthesis. *Journal of Bacteriology* 163: 476–486.

Song W, Joo M, Yeom J-H, Shin E, Lee M, Choi H-K, *et al.* (2019). Divergent rRNAs as regulators of gene expression at the ribosome level. *Nature Microbiology* 4: 515–526.

Spohn M, Klaus K, Wanek W, Richter A. (2016). Microbial carbon use efficiency and biomass turnover times depending on soil depth - Implications for carbon cycling. *Soil Biology and Biochemistry* 96: 74–81.

Srivastava AK, Schlessinger D. (1990). Mechanism and Regulation of Bacterial Ribosomal RNA Processing. *Annu Rev Microbiol* 44: 105–129.

Stone MM, DeForest JL, Plante AF. (2014). Changes in extracellular enzyme activity and microbial community structure with soil depth at the Luquillo Critical Zone Observatory. *Soil Biology and Biochemistry* 75: 237–247.

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23: 1282–1288.

Tas N, Prestat E, McFarland JW, Wickland KP, Knight R, Berhe AA, *et al.* (2014). Impact of fire on active layer and permafrost microbial communities and metagenomes in an upland Alaskan boreal forest. *ISME Journal* 8: 1904–1919.

Team RC. (2018). R: A language and environment for statistical computing. *httpswwwR-projectorg*.

Turner S, Mikutta R, Meyer-Stüve S, Guggenberger G, Schaarschmidt F, Lazar CS, *et al.* (2017). Microbial community dynamics in soil depth profiles over 120,000 years of ecosystem development. *Front Microbiol* 8: 1–17.

Vandekerckhove TTM, Willems A, Gillis M, Coomans A. (2000). Occurrence of novel verrucomicrobial species, endosymbiotic and associated with parthenogenesis in *Xiphinema americanum*- group species (Nematoda, Longidoridae). *International Journal of Systemic and Evolutionary Microbiology* 50: 2197–2205.

VanInsberghe D, Maas KR, Cardenas E, Strachan CR, Hallam SJ, Mohn WW. (2015). Non-symbiotic Bradyrhizobium ecotypes dominate North American forest soils. 9: 2435–2441.

Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, *et al.* (2015). Microbial species delineation using whole genome sequences. *Nucleic Acids Research* 43: 6761–6771.

Vartoukian SR, Palmer RM, Wade WG. (2010). Strategies for culture of 'unculturable' bacteria. *FEMS Microbiology Letters* 309: 1–7.

Větrovský T, Baldrian P. (2013). The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses Neufeld J (ed). *PLoS ONE* 8: e57923–10.

Vieira-Silva S, Rocha E. (2009). The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics. *PLOS Genetics* 6: 1–15.

Vos M, Wolf AB, Jennings SJ, Kowalchuk GA. (2013). Micro-scale determinants of bacterial diversity in soil. *FEMS Microbiol Rev* 37: 936–954.

Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology* 73: 5261–5267.

White RA, Bottos EM, Roy Chowdhury T, Zucker JD, Brislawn CJ, Nicora CD, *et al.* (2016). Moleculo Long-Read Sequencing Facilitates Assembly and Genomic Binning from Complex Soil Metagenomes. Langille M (ed). *mSystems* 1: 309–15.

Wickham H. (2009). *ggplot2: Elegant Graphics for Data Analysis (Use R)*. Springer.

Wilcon HR, Yu D, Peters HK III, Zhou J-G, Court DL. (2002). The global regulator RNase III modulates translation repression by the transcription elongation factor N. *EMBO* 21: 4154–4161.

Will C, Thürmer A, Wollherr A, Nacke H, Herold N, Schrumpf M, *et al.* (2010). Horizon-specific bacterial community composition of German grassland soils, as revealed by pyrosequencing-based analysis of 16S rRNA genes. *Applied and Environmental Microbiology* 76: 6751–6759.

Wilson WA, Roach PJ, Montero M, Baroja-Fernández E, Muñoz FJ, Eydallin G, *et al.* (2010). Regulation of glycogen metabolism in yeast and bacteria. *FEMS Microbiol Rev* 34: 952–985.

Woodcroft BJ, Singleton CM, Boyd JA, Evans PN, Emerson JB, Zayed AAF, *et al.* (2018). Genome-centric view of carbon processing in thawing permafrost. *Nature* 560: 49–54.

Wu Y-W, Simmons BA, Singer SW. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32: 605–607.

Yabe S, Aiba Y, Sakai Y, Hazaka M, Yokota A. (2011). Thermogemmatispora onikobensis gen. nov., sp. nov. and Thermogemmatispora foliorum sp. nov., isolated from fallen leaves on geothermal soils, and description of Thermogemmatisporaceae fam. nov. and Thermogemmatisporales ord. nov. within the class Ktedonobacteria. *International Journal of Systemic and Evolutionary Microbiology* 61: 903–910.

Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. (2016). ggtree: an rpackage for visualization and annotation of phylogenetic trees with their covariates and other associated data McInerny G (ed). *Methods Ecol Evol* 8: 28–36.

Yuan C, Lei J, Cole J, Sun Y. (2015). Reconstructing 16S rRNA genes in metagenomic data. *Bioinformatics* 31: i35–i43.

Yuan H, Ge T, Chen C, O'Donnell AG, Wu J. (2012). Significant Role for Microbial Autotrophy in the Sequestration of Soil Carbon. *Applied and Environmental Microbiology* 78: 2328–2336.

Zeng YH, Koblížek M, Li YX, Liu YP, Feng FY, Ji JD, *et al.* (2012). Long PCR-RFLP of 16S-ITS-23S rRNA genes: a high-resolution molecular tool for bacterial genotyping. *J Appl Microbiol* 114: 433–447.

Zhi X-Y, Zhao W, Li W-J, Zhao G-P. (2012). Prokaryotic systematics in the genomics era. *Antonie van Leeuwenhoek* 101: 21–34.

Zhou J, Xia B, Huang H, Treves DS, Hauser LJ, Mural RJ, *et al.* (2003). Bacterial phylogenetic diversity and a novel candidate division of two humid region, sandy surface soils. *Soil Biology and Biochemistry* 35: 915–924.

APPENDIX

CHAPTER II APPENDIX
UNCULTURED OLIGOTROPHIC MICROBES THRIVE IN SUBSURFACE SOILS

Materials and Methods

Sample collection and processing

Samples were collected from the network of 10 Critical Zone Observatories (CZOs, http://criticalzone.org) across the US: Southern Sierra (CA), Boulder Creek (CO), Reynolds (ID), Shale Hills (PA), Calhoun (SC), Luquillo (PR), Intensively Managed Landscapes (IL/IA/MN), Catalina/Jemez (AZ/NM), Eel River (CA), and Christina River (DE/PA). Volunteers from each CZO excavated two separate soil profiles ("sites") selected to represent distinct soil types and landscape positions. Soils were collected at peak greenness as estimated from the Normalized Difference Vegetation Index and Enhanced Vegetation Index measured by NASA's MODIS (MODerate-resolution Imaging Spectroradiometer) instrument. These collections were conducted between April 2016 and November 2016, with the exception of the Eel River CZO samples, which were collected in May 2017. Volunteers were asked to sample in 10-cm increments to a depth of at least 100 cm, or to refusal. It was not possible to reach 100 cm at all sites. Because few sites were sampled past 90-100 cm, samples from deeper than 100 cm were not used in this study. For all but two sites, soils from the pit or core were collected aseptically using either a soil knife or a coring auger inserted into the pit wall horizontally, integrating soil from each 10-cm increment.

All soil samples were sent to the University of California, Riverside for processing. A portion of each field sample was sieved (< 2 mm, ASTM No. 10), homogenized, and divided into subsamples for further analyses, with subsamples stored at either 4°C, −20°C, or −80°C. For some soils (particularly some wet, finely textured depth intervals), sieving was not practical under field-moist conditions. These samples were homogenized by mixing, but during the subsampling process, larger root and rock fragments were removed by hand. In addition, as samples from Shale Hills site 2 (70—100 cm depth) consisted almost entirely of medium-sized rocks, soil was collected by manually crushing rocks with a ceramic mortar and pestle; this material was then passed through a 2-mm sieve.

DNA was extracted from subsamples frozen at −20°C using the DNeasy PowerLyzer

PowerSoil kit (Qiagen, Germantown, MD, USA), according to the manufacturer's instructions with minor modifications to increase yield and final DNA concentration based on the assumption that some sites and depths would have a relatively low microbial biomass. Specifically, 0.25 g of soil was weighed in triplicate (i.e., three 0.25 g aliquots = 0.75 g total soil per sample) from one frozen aliquot of sieved soil (from the subsample reserved specifically for DNA extractions). Extractions on each 0.25 replicate aliquot proceeded in parallel, until the stage when DNA was eluted onto the spin filter; replicates were pooled at this point onto a single filter, and extractions proceeded from this point as a single sample. In addition, the final step of elution of the DNA from the filter was conducted with 50 µL of elution buffer, instead of 100 µL; the initial flow-through was reapplied to the filter and passed through a second time to further increase yield.

Soil characteristics

Frozen subsamples (stored at −20°C) were shipped to the University of Illinois at Urbana-Champaign for characterization of soil physicochemical properties. Soil C and N concentrations were measured on freeze-dried, sieved, and ground subsamples using a Vario Micro Cube elemental analyzer (Elementar, Hanau, Germany). Approximately 1 g of each subsample was also extracted in 30 mL of 0.5 N HCl for determination of Fe(III) and Fe(II) concentrations using a modified ferrozine assay (Liptzin and Silver, 2009). Soil texture was measured on oven-dried and sieved soil following Gee and Bauder (1986).

Soil pH and gravimetric water content were measured using modified Long Term Ecological Research (LTER) protocols, as per Robertson et al. (1999). Soil pH was determined using 15 g of field-wet soil and 15 mL of Milli-Q water (Millipore Sigma, Burlington, Massachusetts), and was measured on a Hannah Instruments (Woonsocket, RI) HI 3220 pH meter with a HI 1053B pH electrode, designed for use with semi-solids. For determining gravimetric water content, we oven-dried 7 g of soil at 105 ° C, for a minimum of 24 hours.

Amplicon-based 16S rRNA gene analyses

To characterize the bacterial and archaeal communities in each sample, we used the barcoded primer pair 515f/806r for sequencing the V4-V5 region of the 16S rRNA gene following methods described previously (Leff et al., 2015). We amplified this gene region in triplicate reactions per sample, combined these products, and normalized the concentration of each sample to 25 ng using SequalPrep Normalization Plate Kits (Thermo Fisher Scientific, Waltham, MA). All samples were then pooled and sequenced on the Illumina MiSeq (2x150 paired end chemistry) at the University of Colorado Next-Generation Sequencing Facility. The sample pool included several kit controls and no template controls to check for possible contamination.

Sequences were processed using a combination of QIIME and USEARCH commands to demultiplex, quality-filter, remove singletons, and merge paired end reads. Sequences were classified into exact sequence variants (ESVs) using UNOISE2 (Edgar, 2016) with default settings and taxonomy was assigned against the Greengenes 13_8 database (DeSantis et al., 2006) using the RDP classifier (Cole et al., 2013). ESVs with greater than 1% average abundance across all sequenced controls were classified as contaminants and removed from further analyses, along with ESVs identified as mitochondria and chloroplast. The entire dataset was then rarefied to 3400 sequences per sample.

Shotgun metagenomic analyses

One soil profile from each CZO was selected for shotgun sequencing - we chose the sites that exhibited the most dissimilarity in microbial community composition through the soil profile, as we were interested in changes most associated with soil depth. Using the same DNA as used for the amplicon sequencing effort, we generated metagenomic libraries using the TruSeq DNA LT library preparation kit (Illumina, San Diego, CA). All samples were pooled and sequenced on an Illumina NextSeq run using 2x150bp paired end chemistry at the University of Colorado Next-Generation Sequencing Facility. Prior to downstream analysis, we merged and quality filtered the paired-end metagenomic reads with USEARCH. After quality filtering we had an average of 8.8 million quality-filtered reads per sample (range = 1.9 -15.4 million reads, we only included samples with at last 1 million reads). These sequences were uploaded to MG-RAST (Meyer et al., 2008) for annotation. We used Metaxa2 (Bengtsson-Palme et al., 2015) with default settings to analyze all microbial communities (bacterial, archaeal, and eukaryotic) in each sample. All statistical analyses were done in R studio and all figures were created with ggplot2 (Wickham, 2009).

Assembly, annotation, and characterization of AD3 genomes

We assembled two genomes belonging to the candidate phylum AD3 (Ji et al., 2017) from individual metagenomes obtained from Calhoun site 1 (60-70cm) and Shale Hills site 1 (90-100 cm). These two soil samples were selected for deeper sequencing based on the high abundance of the phylum AD3 (~60% of amplicon 16S rRNA gene reads at Calhoun, ~23% at Shale Hills). This sequencing effort yielded 57.7 million paired-end reads for Calhoun 60-70cm and 65.6 million paired end reads for Shale Hills 90-100cm.

Genomes were assembled using unpaired reads that had been filtered using sickle (-q 20 -l 50). We used Megahit (Li et al., 2015) with the bulk preset to build the assembly, and binned the assembly with MaxBin 2.2.1 (Wu et al., 2016) . We used a script that cycled through MaxBin conditions (-min_contig_length 1100 -1500 and -prob_threshold 0.95 - 0.99) and used checkM (Parks et al., 2015) to pick the best bins. Bins were then manually curated using a combination of scaffold abundance, tetranucleotide frequency, and GC content. After selecting the highest quality bins from each sample, we ran Metaxa2 on the bins themselves to detect SSU or LSU rRNA genes that could be used to determine taxonomic affiliations. To double check that each bin was affiliated with the AD3 candidate phylum, we used the concatenated marker gene phylogeny generated from checkM to compare the placement of our genomes to three previously published AD3 genomes (Ji et al., 2017). Our AD3 genomes clustered with the genomes from Ji et al. (Ji et al., 2017) and fell near the Chloroflexi and Armatimonadetes on the tree. Both genomes were submitted to IMG for annotation under the taxon IDs 2756170100 and 2767802471. Based on checkM estimates, both genomes are substantially complete with medium to high contamination (bin JG37: 74.54% complete, 12.66% contamination; bin3: 72.65% complete, 10.68% contamination) See Supplemental Table S2.1 for additional genome details.

Phylogenetic tree of CoxL genes

The evolutionary history of the AD3 coxL genes was inferred by using the Maximum Likelihood method based on the JTT matrix-based model (Jones et al., 1992). The tree with the highest log likelihood (-24038.8804) is shown. The percentage of trees in which the associated taxa clustered together is shown below the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. A discrete gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 1.10)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 6.75% sites). The resulting tree was drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 67 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 526 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016) and the tree was plotted in ggtree (Yu et al., 2016).

Calculation of maximum growth rate proxy ΔENC'

Because there are no cultivated members of phylum AD3, we calculated ΔENC' to estimate potential growth rates as described previously (Novembre, 2002; Vieira-Silva and Rocha, 2009). We also calculated ΔENC' on complete genomes in NCBI that matched amplicon sequences in our dataset with >=99% sequence similarity. We used this set of genomes to represent the bacteria, or at least closely related lineages of bacteria, found in the same soil profiles studied here to establish a range for microbial growth rates in soil. We ran ENCprime (Novembre, 2002) with default options on both concatenated ribosomal protein sequences and concatenated genome sequences, and calculated ΔENC' as described in Vieira-Silva (2009).

Spore selection treatment

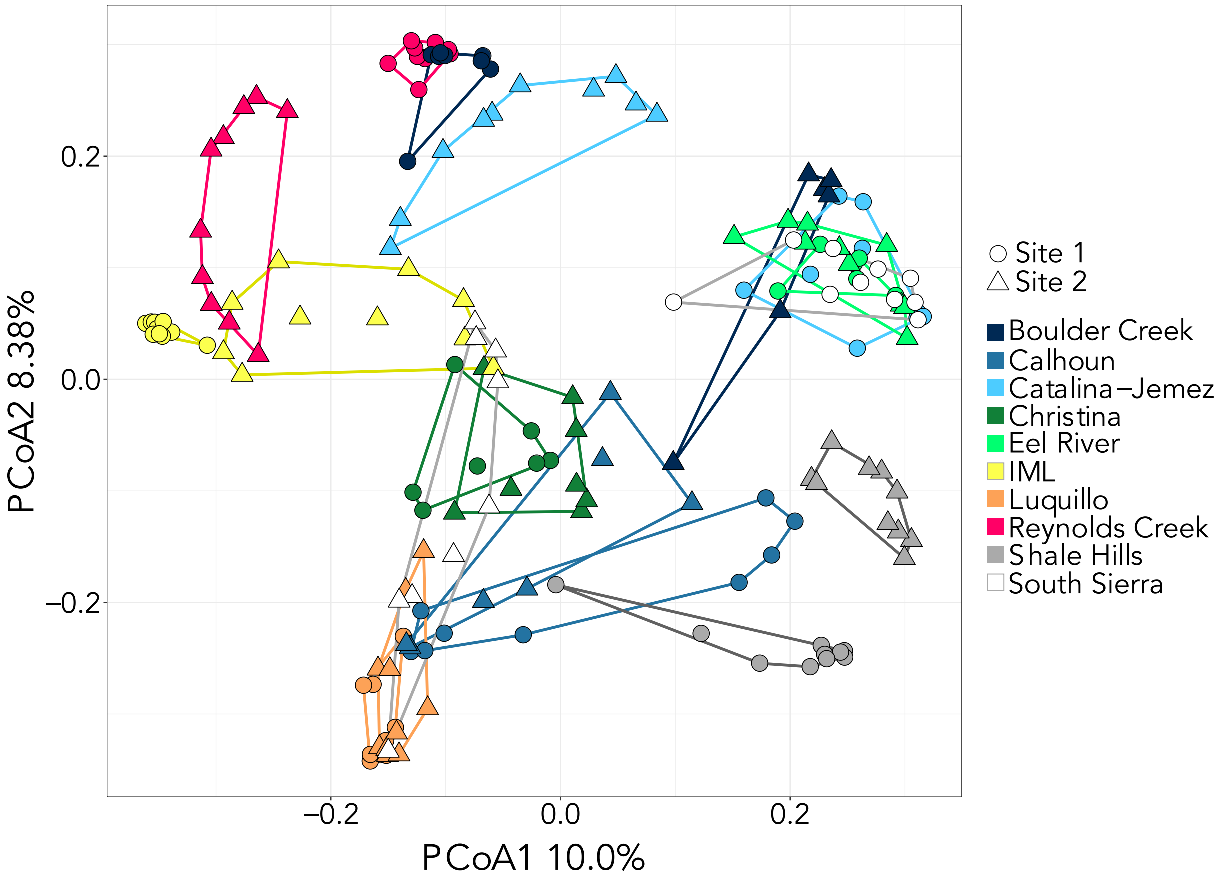
When we examined our AD3 genomes, we found numerous genes linked to spore formation. Therefore, we adapted a method previously used in human stool samples (Browne et al., 2016) to select for spores in a culture independent manner in three soil samples from our study (Calhoun site 1, soils 50-60cm, 60-70cm, and Calhoun site 2, soil 50-60cm). To select for spores, we incubated 0.04g of each soil in 70% ethanol for 4 hours under constant agitation with the goal of killing vegetative cells. In our control samples, we performed the same incubation with phosphate-buffered saline (PBS). After the incubations, we washed both sets of samples with PBS three times, then applied propidium monoazide (PMA) to the ethanol-treated samples as described previously (Carini et al., 2016). We used PMA to block the amplification of DNA from cells with compromised membranes, ensuring that only those cells capable of surviving the harsh ethanol treatment would be amplified in subsequent PCRs. We PCR amplified, sequenced, and processed these samples as previously described. We restricted our analysis to the top 1000 most abundant phylotypes to remove rare taxa and used the Wilcoxon test to identify enriched taxa, scoring taxa as "possible spore formers" if they had False Discovery Rate (FDR) corrected p-values greater than 0.05. These taxa are presented in Supplemental Table S2.3.

Data availability

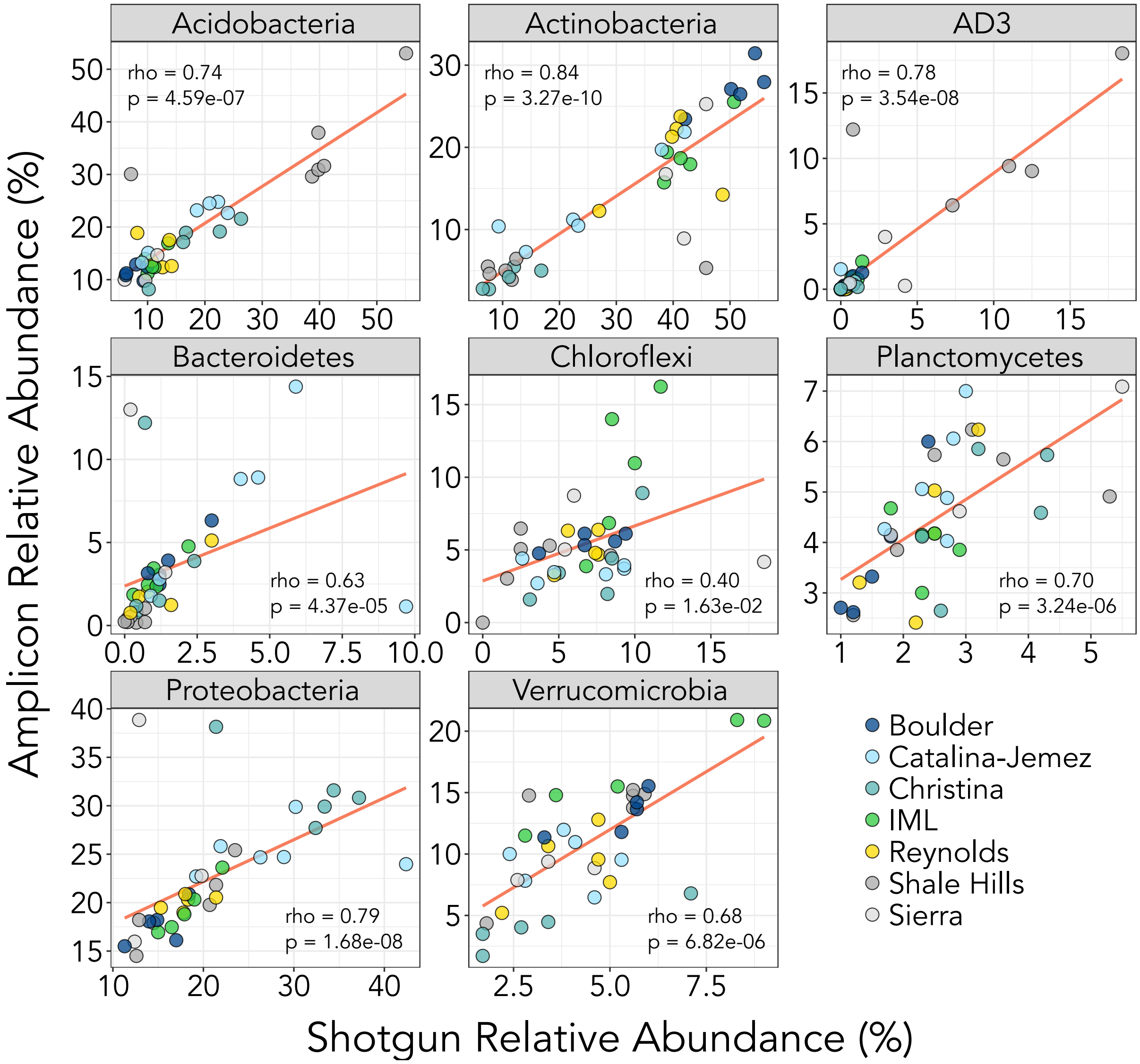
Both AD3 genomes are publicly available on IMG under taxon IDs 2756170100 and 2767802471. The merged, quality filtered, and unassembled shotgun sequences are available under MG-RAST project ID mgp80869. The raw, unmerged 16S amplicon sequences are available on figshare at 10.6084/m9.figshare.4702711.

Acknowledgements

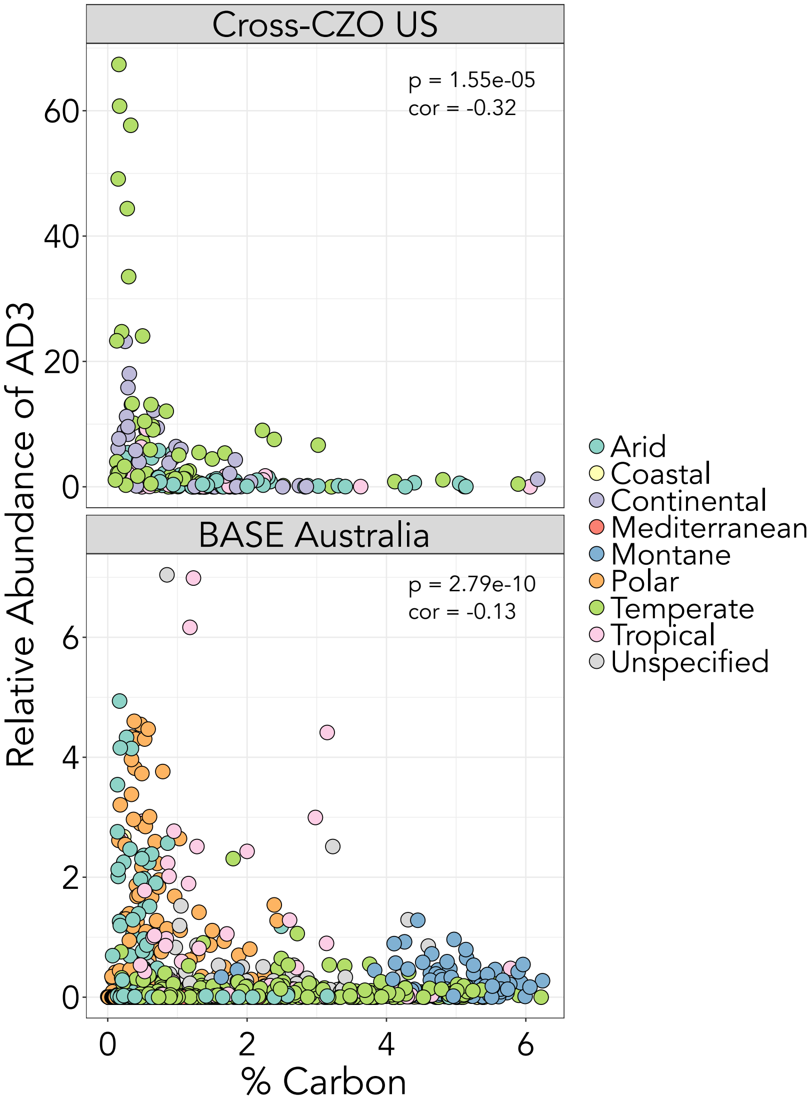
We want to thank Chelsea Carey, Neal Blair, and Andrew Bissett for their contributions to this research effort and Belinda Ferrari for providing feedback on a previous draft of this manuscript. Funding for this work was provided the U.S. National Science Foundation's Critical Zone Observatories program with additional funding provided by the NSF EarthCube program (ICER-1541047) and the Macrosystems program (EF-1550920).



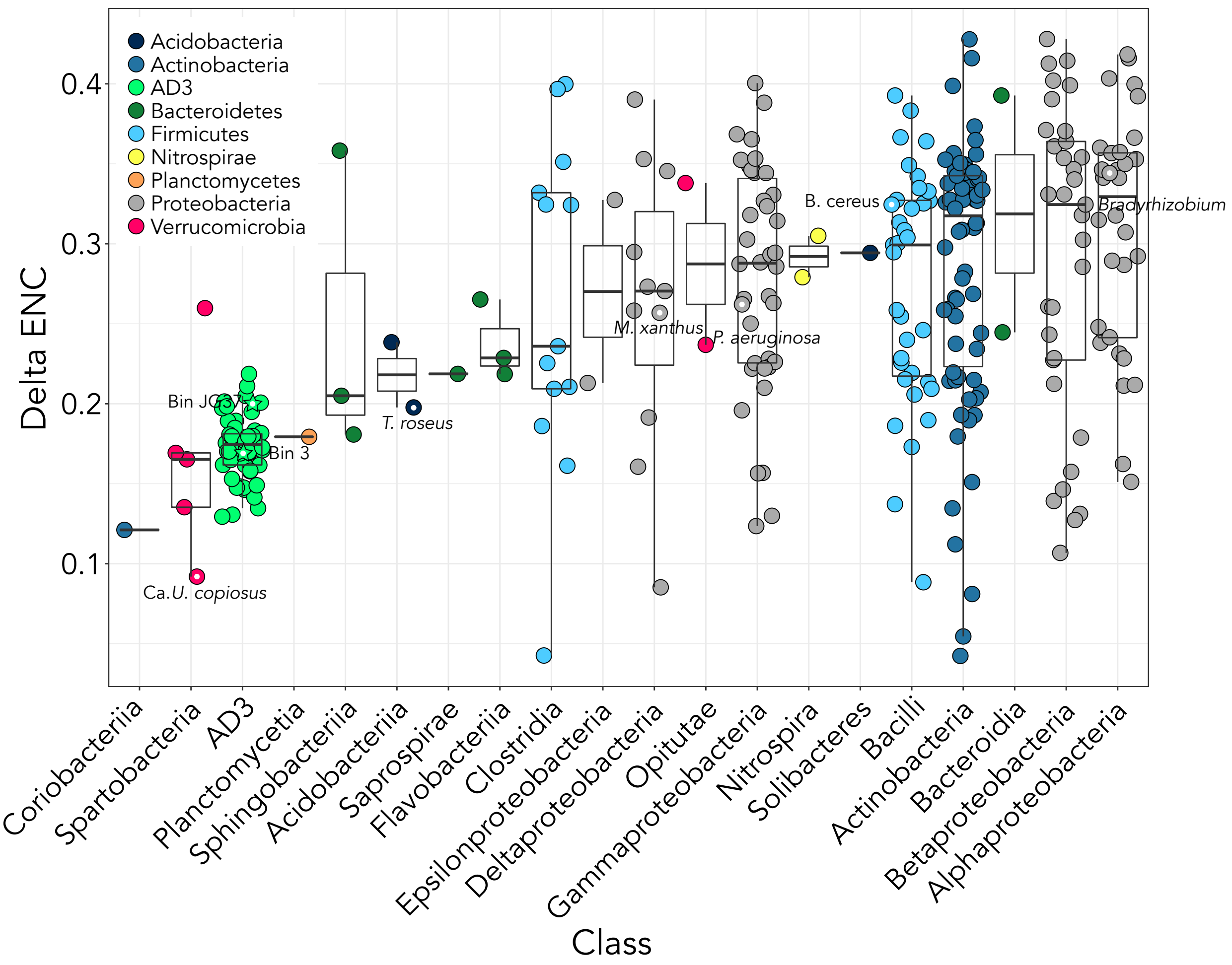
Supplemental Figure S2.1: Ordination plot showing differences in overall microbial community composition across the 20 sampled profiles (two per Critical Zone Observatory). The principal coordinates analysis is based on Bray-Curtis dissimilarities calculated from the 16S rRNA gene sequencing effort (amplicon data). This ordination plot shows that the differences in communities between profiles are typically larger than the differences in communities across different depths within individual profile, a conclusion supported by the associated PerMANOVA analyses.



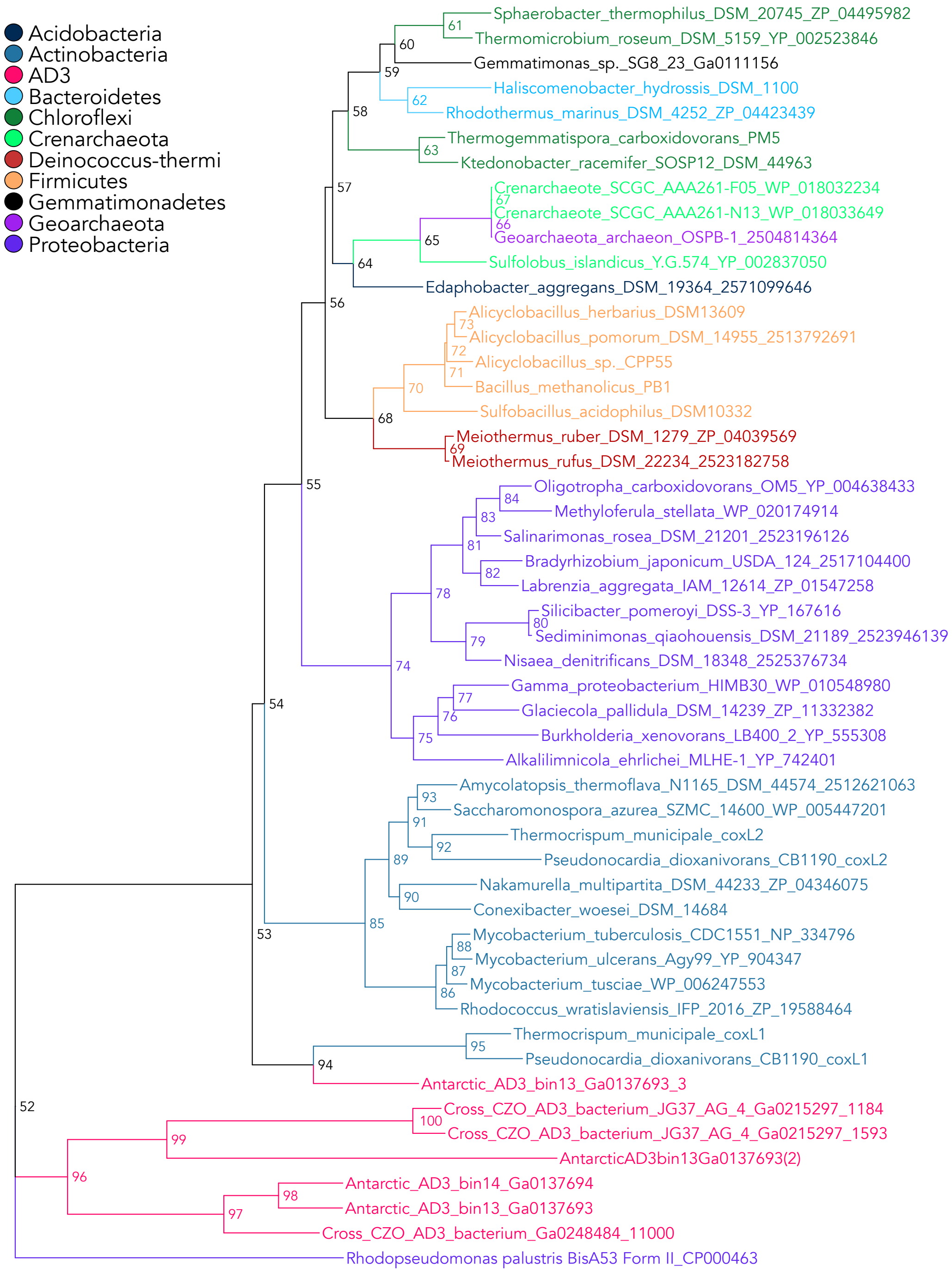
Supplemental Figure S2.2: Relative abundances of the eight most abundant bacterial phyla in our dataset are well correlated between 16S rRNA gene amplicon and shotgun metagenomic methods. We used Metaxa2 (Bengtsson-Palme et al., 2015) to search for SSU rRNA gene fragments in our metagenomic data. P-values and rho values indicate Spearman correlations.



Supplemental Figure S2.3: The relative abundance of phylum AD3 is negatively correlated with soil carbon concentrations across two independent soil datasets. The top panel features data from this study (20 soil profiles across the U.S., 177 soils in total) while the bottom panel draws from a dataset encompassing 1006 surface soils (0-10 cm depth) collected from across Australia as part the BASE project (Biomes of Australian Soil Environments; Bissett et al., 2016). P-values and correlation coefficients (cor) are from Pearson correlations.



Supplemental Figure S2.4: Members of the phylum AD3 are predicted to have low maximum potential growth rates based on the growth rate proxy ΔENC' (a metric of codon usage bias). ΔENC' ranges from our AD3 genomes, the Antarctic AD3 genomes (Ji et al., 2017), the thawing permafrost AD3 genomes (Woodcroft et al., 2018), and a set of genomes matched to our 16S rRNA gene amplicon sequences are shown, arranged by taxonomic affiliation. ΔENC' is positively correlated with growth rate in bacterial and archaeal genomes(Vieira-Silva and Rocha, 2009). Select genomes are labeled for additional context and indicated with a central white dot.


● Acidobacteria
● Actinobacteria
● AD3
● Bacteroidetes
● Chloroflexi
● Crenarchaeota
● Deinococcus-thermi
● Firmicutes
● Gemmatimonadetes
● Geoarchaeota
● Proteobacteria
Sphaerobacter_thermophilus_DSM_20745_ZP_04495982
61
Thermomicrobium_roseum_DSM_5159_YP_002523846
60
Gemmatimonas_sp._SG8_23_Ga0111156
59
Haliscomenobacter_hydrossis_DSM_1100
62
Rhodothermus_marinus_DSM_4252_ZP_04423439
58
Thermogemmatispora_carboxidovorans_PM5
63
Ktedonobacter_racemifer_SOSP12_DSM_44963
57
Crenarchaeote_SCGC_AAA261-F05_WP_018032234
67
Crenarchaeote_SCGC_AAA261-N13_WP_018033649
66
Geoarchaeota_archaeon_OSPB-1_2504814364
65
Sulfolobus_islandicus_Y.G.574_YP_002837050
64
Edaphobacter_aggregans_DSM_19364_2571099646
56
Alicyclobacillus_herbarius_DSM13609
73
Alicyclobacillus_pomorum_DSM_14955_2513792691
72
Alicyclobacillus_sp._CPP55
71
Bacillus_methanolicus_PB1
70
Sulfobacillus_acidophilus_DSM10332
68
Meiothermus_ruber_DSM_1279_ZP_04039569
69
Meiothermus_rufus_DSM_22234_2523182758
55
Oligotropha_carboxidovorans_OM5_YP_004638433
84
Methyloferula_stellata_WP_020174914
83
Salinarimonas_rosea_DSM_21201_2523196126
81
Bradyrhizobium_japonicum_USDA_124_2517104400
82
Labrenzia_aggregata_IAM_12614_ZP_01547258
78
Silicibacter_pomeroyi_DSS-3_YP_167616
80
Sediminimonas_qiaohouensis_DSM_21189_2523946139
79
Nisaea_denitrificans_DSM_18348_2525376734
74
Gamma_proteobacterium_HIMB30_WP_010548980
77
Glaciecola_pallidula_DSM_14239_ZP_11332382
76
Burkholderia_xenovorans_LB400_2_YP_555308
75
Alkalilimnicola_ehrlichei_MLHE-1_YP_742401
54
Amycolatopsis_thermoflava_N1165_DSM_44574_2512621063
93
Saccharomonospora_azurea_SZMC_14600_WP_005447201
91
Thermocrispum_municipale_coxL2
92
Pseudonocardia_dioxanivorans_CB1190_coxL2
89
Nakamurella_multipartita_DSM_44233_ZP_04346075
90
Conexibacter_woesei_DSM_14684
53
85
Mycobacterium_tuberculosis_CDC1551_NP_334796
88
Mycobacterium_ulcerans_Agy99_YP_904347
87
Mycobacterium_tusciae_WP_006247553
86
Rhodococcus_wratislaviensis_IFP_2016_ZP_19588464
Thermocrispum_municipale_coxL1
95
Pseudonocardia_dioxanivorans_CB1190_coxL1
94
Antarctic_AD3_bin13_Ga0137693_3
52
Cross_CZO_AD3_bacterium_JG37_AG_4_Ga0215297_1184
100
Cross_CZO_AD3_bacterium_JG37_AG_4_Ga0215297_1593
99
AntarcticAD3bin13Ga0137693(2)
96
Antarctic_AD3_bin14_Ga0137694
98
Antarctic_AD3_bin13_Ga0137693
97
Cross_CZO_AD3_bacterium_Ga0248484_11000
Rhodopseudomonas palustris BisA53 Form II_CP000463

Supplemental Figure S2.5: AD3 genomes assembled from cross-CZO soil metagenomes contain form II CO dehydrogenases (coxL). Form II coxL genes may not be associated with the ability to oxidize carbon monoxide (King and Weber, 2007). However the Antarctic bin13 AD3 genome (Ji et al., 2017) contains a form I coxL sequence - indicating that at least some members of this phylum are likely capable of CO oxidation.

Supplemental Table S2.1: Genome Statistics

AD3 Genomes from this study						
IMG_ID	Name	% Complete	% Contam.	Contigs	Assembled length	GC%
2756170100	bin JG37	74.54	12.66	661	2990252 bp	67%
2767802471	bin 3	72.65	10.68	1779	3428017 bp	61%

AD3 Genomes from Ji et al, 2017

2698536734	bin 12	95.32	2.31	302	2961190 bp	69%
2698536735	bin 13	96.3	0	71	2960892 bp	67%
2698536736	bin 14	92.44	4.63	358	5287456 bp	68%

Supplemental Table S2.2: Spore forming genes from AD3 genomes and closest spore-forming relative *K. racemifer*

Stage	Gene	Bin JG-37	Bin 3	*K. racemifer*
stage 0	spo0A	0	0	0
stage 0	sigH (spo0H)	0	0	0
stage 0	spo0J	2	2	2
stage 0	obgE	1	0	1
stage 0	spo0F	0	0	1
stage II	spoIIAA	0	0	2
stage II	spoIIAB	0	0	0
stage II	sigF	0	0	0
stage II	spoIID	2	0	0
stage II	spoIIE (spoIIH)	0	0	0
stage II	spoIIGA	0	0	0
stage II	sigE	0	0	0
stage II	spoIIM	1	0	0
stage II	spoIIP	0	0	0
stage II	spoIIR	0	0	0
stage III-IV	cwlD	0	0	0
stage III-IV	dacB	0	0	0
stage III-IV	spoIIIE	2	1	4
stage III-IV	spoIIIJ	2	1	3
stage III-IV	spoIIIAA	0	1	1
stage III-IV	spoVS	1	1	0
stage III-IV	spoVD	1	1	0
stage III-IV	spoVK	0	0	2
stage III-IV	spoVC	0	1	1
stage III-IV	spoVR	0	0	1
stage III-IV	spoIVFB	4	1	3
stage III-IV	spoIVCA	0	0	6
spore coat	spoIVA	0	0	0
spore coat	alr (yncD)	1	1	1
spore coat	CotA	0	1	1
spore coat	CotJC	0	0	3
germination	gpr	0	0	0
germination	lgt (gerf)	0	0	0
germination	YaaH	1	1	2
germination	CgeB	1	0	0
germination	YhbH	1	1	2
Total	-	20	13	36

Supplemental Table S2.3: Taxa enriched through spore-selection

Taxonomy of ESVs significantly enriched in spore selection treatment	confirmed spore former?
Pseudomonas veronii	plant-associated
Pseudomonas fragi	no
f_Myxococcaceae;g_Anaeromyxobacter	yes
o_Myxococcales;f_Myxococcaceae	yes
c_Deltaproteobacteria;o_Myxococcales	yes
o_Rhodospirillales;f_Rhodospirillaceae	no
o_Rhizobiales;f_Methylocystaceae	no
f_Hyphomicrobiaceae;g_Rhodoplanes	no
f_Bradyrhizobiaceae;g_Bradyrhizobium	plant-associated
c_Alphaproteobacteria;o_Rhizobiales	no
o_Gemmatales;f_Gemmataceae	no
o_Nitrospirales;f_0319-6A21	uncultivated
p_Gemmatimonadetes;c_Gemm-1	uncultivated
p_GAL15;c_	uncultivated
p_Firmicutes;c_Bacilli	yes
o_Bacillales;f_Paenibacillaceae	yes
p_Chloroflexi;c_TK10;o_B07_WMSP1;f_FFCH4570	yes
p_Chloroflexi;c_TK10;o_B07_WMSP1	yes
o_Thermogemmatisporales;f_Thermogemmatisporaceae	yes
Sphingobacterium faecium	no
f_Flavobacteriaceae;g_Flavobacterium	no
p_AD3;c_JG37-AG-4	uncultivated
p_AD3;c_ABS-6	uncultivated
f_Nocardiaceae;g_Rhodococcus;s_	yes
o_Actinomycetales;f_Micromonosporaceae	yes
Arthrobacter psychrolactophilus	yes
o_Actinomycetales;f_Micrococcaceae	yes
o_Actinomycetales;f_Actinosynnemataceae	yes
c_Actinobacteria;o_Actinomycetales	yes
p_Acidobacteria;c_TM1	uncultivated
c_DA052;o_Ellin6513	uncultivated
f_Koribacteraceae;g_Candidatus Koribacter;s_	no
o_Acidobacteriales;f_Koribacteraceae	no
c_Acidobacteria-6;o_iii1-15	uncultivated
p_Acidobacteria;c_Chloracidobacteria	uncultivated
p_Acidobacteria;c_Chloracidobacteria	uncultivated
o_E2;f_Methanomassiliicoccaceae	no

CHAPTER III APPENDIX
GENOME REDUCTION IN AN ABUNDANT AND UBIQUITOUS SOIL BACTERIUM 'CANDIDATUS UDAEOBACTER COPIOSUS'

Materials and Methods

Estimating the abundances and distributions of Verrucomicrobia in soil

While five abundant Verrucomicrobia phylotypes were described in Fierer et al. 2013, a single phylotype with 99% identity to the clone DA101 (Felske and Akkermans, 1998) was clearly dominant. We searched previously published soil datasets for representative sequences with 100% identity to this DA101 phylotype, including 31 soils from United States native tallgrass prairies (Fierer et al., 2013), 64 soils from matched forest and grassland sites across North America (Crowther et al., 2014), 595 soils collected from Central Park in New York City (Ramirez et al., 2014), 367 grassland soils collected from North America, Europe, Australia, and Africa (Leff et al., 2015), and a cross-biome collection of 15 desert and non-desert soils from across the globe (Fierer et al., 2012). We also included a dataset from a grassland terrace near Boulder, Colorado (105.23W, 40.12N, Table Mountain) where 29 soils were collected from a depth of 25 cm within a 100m^2 area on January 28th, 2015. Collectively these datasets represent 1101 unique soil samples collected from a wide range of ecosystem and soil types.

For all samples, DNA was extracted with the MoBio PowerSoil kit and the V4 region of the 16S rRNA gene was amplified in triplicate with the 515f/806r primer pair. After normalization to equimolar concentrations, amplicons were sequenced on an Illumina MiSeq (151 bp paired end) at the University of Colorado BioFrontiers Institute Next-Gen Sequencing Facility. Sequences were processed as described previously (Leff et al., 2015). In brief, we used a combination of QIIME (Caporaso et al., 2010b) and UPARSE (Edgar, 2013) to quality-filter, remove singletons, and merge paired reads. Sequences were assembled into phylotypes at the 97% identity level using UCLUST (Edgar, 2010a). Taxonomy was assigned using the Greengenes 13_8 database (DeSantis et al., 2006) and the Ribosomal Database Project classifier (Wang et al., 2007) and each dataset was rarefied independently.

As PCR primer biases can misestimate the relative abundances of Verrucomicrobia (Bergmann et al., 2011; Guo et al., 2016), we also estimated the abundances of the DA101 phylotype directly from shotgun metagenomic data. We used Metaxa2 with default settings (Bengtsson-Palme et al., 2015) to extract bacterial 16S rRNA gene sequences from shotgun metagenomic data compiled from previous analyses of 75 different soils after rarefaction (Leff et al., 2015; Fierer et al., 2012). Extracted 16S rRNA gene fragments were matched to Greengenes full-length sequences at 99% ID using the usearch7 command usearch_global. The matched Greengenes sequences were then clustered and assigned taxonomy as described above. All statistical tests were carried out in R and ggplot2 was used for all plots unless specifically mentioned. Variances between groups tested were within one order of magnitude.

Describing the phylogenetic diversity of soil Verrucomicrobia

We reconstructed near-full length 16S rRNA gene sequences to build a phylogeny of soil Verrucomicrobia from six soil samples (see Supplemental Table S3.1) that were selected to represent geographically distinct grasslands with a range of verrucomicrobial abundances. We extracted DNA from each of these soils as described previously (Leff et al., 2015) and used the 27f/1392r primer pair to amplify near full-length 16S rRNA genes as described in (Miller et al., 2013). The amplicons were sheared using the Covaris M220 (Covaris, Woburn, MA) and 16S rRNA gene libraries were prepared using TruSeq DNA LT library preparation kits (Illumina, San Diego, CA). Samples were pooled and sequenced on an Illumina MiSeq (2x300bp) at the University of Colorado Next Generation Sequencing Facility.

After quality filtering of sequences, near full length SSU sequences were reconstructed using EMIRGE (Miller et al., 2013). After 40 iterations, sequences were merged into phylotypes with ≥97% similarity. Reconstructed sequences were trimmed to 1200 bp and all sequences were further clustered at 95% identity due to gaps in some assemblies. Full-length 16S rRNA sequences from named verrucomicrobial isolates were aligned along with the reconstructed sequences using PyNAST (Caporaso et al., 2010a). A UPGMA tree was constructed using the R packages seqnir, phangorn, and ape and visualized with GraPhlAn (Asnicar et al., 2015) (R 3.2.2, version 0.9.7).

Assembly and annotation of the dominant soil Verrucomicrobia genome

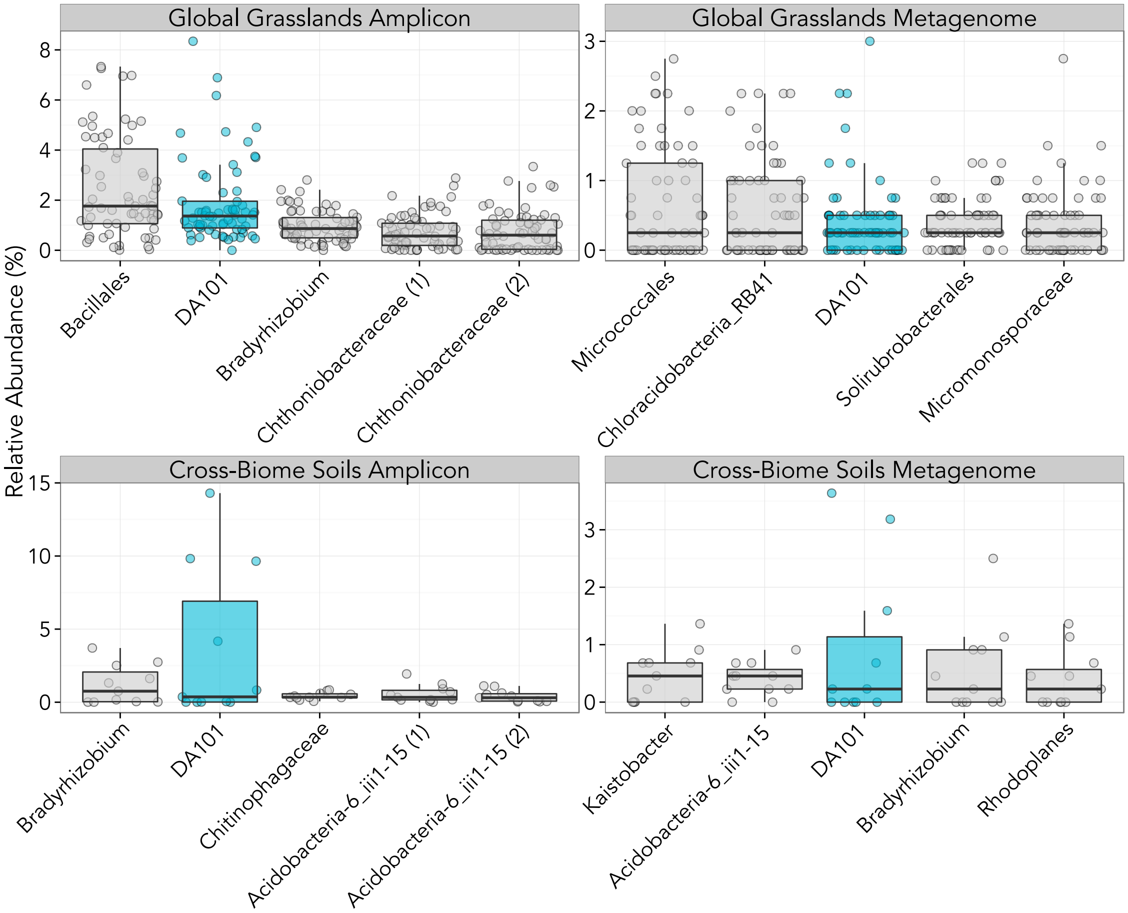
We assembled the genome of '*Candidatus* Udaeobacter copiosus' from a metagenome of a U.S. prairie soil sample (NTP21, Hayden, IA) estimated to have particularly high abundances of bacteria within the DA101 clade (Fierer et al., 2013). Fragmented DNA extracted from this soil was prepared for sequencing using WaferGen's PrepX ILM DNA library Kit (WaferGen Biosystems Inc, Fremont, CA) and the Apollo 324 Automated Library Prep System for library generation. The library was sequenced on one Illumina HiSeq2000 lane (2×101 bp), yielding 17 Gb of sequence with an average paired-end insert size of 345 bp. Low quality reads were trimmed using Sickle v. 1.29 with a quality score threshold of Q=3, or removed if trimmed to <80 bp long (https://github.com/najoshi/sickle). The sequences were assembled using IDBA_ud v. 1.1.0 (Peng et al., 2012) with a kmer range of 40 to 70 and step size of 15. To improve recovery of the most abundant Verrucomicrobia, the genome was selectively re-assembled using Velvet with a kmer size of 59, and expected kmer coverage of 11.5 (range 7.5 to 15.5). To bin contigs ≥2 kb long, genes and protein sequences were predicted using Prodigal v. 2.60 in metagenomics mode (Hyatt et al., 2010). For each contig, we determined the GC content, coverage, and the phylogenetic affiliation based on the best hit for each predicted protein in the Uniref90 database (Suzek et al., 2007) (Sept-2013) following ublast searches. We also constructed emergent self-organizing maps (ESOM; Dick et al., 2009) using tetranucleotide frequencies of 5 kb DNA fragments. A combination of these approaches was used to identify the genome. The draft genome was uploaded to IMG for annotation under the taxon ID 2651869889.

We estimate that the *Ca*. U. copiosus genome is approximately 94% complete, based on domain-specific single copy housekeeping genes commonly used to estimate genome completion (Finn et al., 2016). This list of single copy genes has been used to estimate genome completeness in several recent studies(Herlemann et al., 2013; Anantharaman et al., 2016). When we analyzed the genome using another metric of genome completeness (checkM; Parks et al., 2015), the results suggested that the genome was 80% complete with 4% contamination, a level categorized as a 'substantially complete draft with low contamination'. This level of completeness is similar to several other recent genomes assembled from metagenomes (Baker et al., 2016; Garcia et al., 2015). However, because checkM relies on lineage-specific marker genes, the completeness of genomes without lineage representation can often be underestimated (Parks et al., 2015). As there is only one complete genome for the entire class Spartobacteria (*C. flavus*), the checkM genome completeness estimate for *Ca*. U. copiosus may likewise be underestimated. Simply put, there are limitations and caveats associated with any genome completeness measure and the true completeness of the *Ca*. U. copiosus genome likely lies somewhere between these estimates.

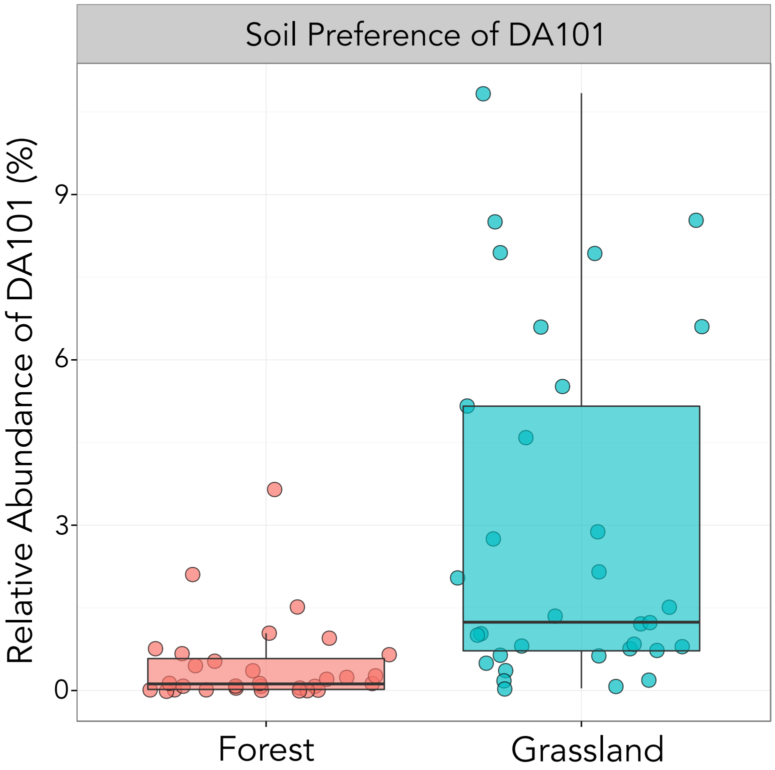
No rRNA genes were annotated by IMG, so we used Metaxa2 with default settings on the unassembled sequences to extract any 16S rRNA genes. Metaxa2 recovered two ~500¬ bp 16S rRNA gene fragments at 23-29× coverage which aligned to separate regions of the full-length 16S rRNA gene from the closest related verrucomicrobial genome (*C. flavus*). Because these two rRNA gene fragments have the same coverage as the genome (27x) and align to separate regions of one 16S rRNA gene, it is likely that *Ca.* U. copiousus encodes a single rRNA operon, similar to its closest relative *C. flavus* (Sangwan et al., 2004; Kant et al., 2011) and all other sequenced heterotrophic soil Verrucomicrobia (Supplemental Table 4).

Data Availability: The draft genome of '*Candidatus* Udaeobacter copiosus' is publicly available in the Integrated Microbial Genomes (IMG) database under the IMG genome ID 2651869889. Raw sequences from which the *Ca.* U. copiosus genome was assembled are available at the Sequence Read Archive (SRA) under the bioproject ID PRJNA342239. Amplicon sequences and associated metadata generated exclusively for this study are available at figshare at dx.doi.org/10.6084/m9.figshare.3363505.v3. Accession numbers for all other amplicon datasets have been previously published. The raw sequences used for EMIRGE near full-length 16S amplicon reconstruction are also available at figshare at dx.doi.org/10.6084/m9.figshare.3799422.v1. All other datasets supporting these findings are available from the corresponding author upon request.

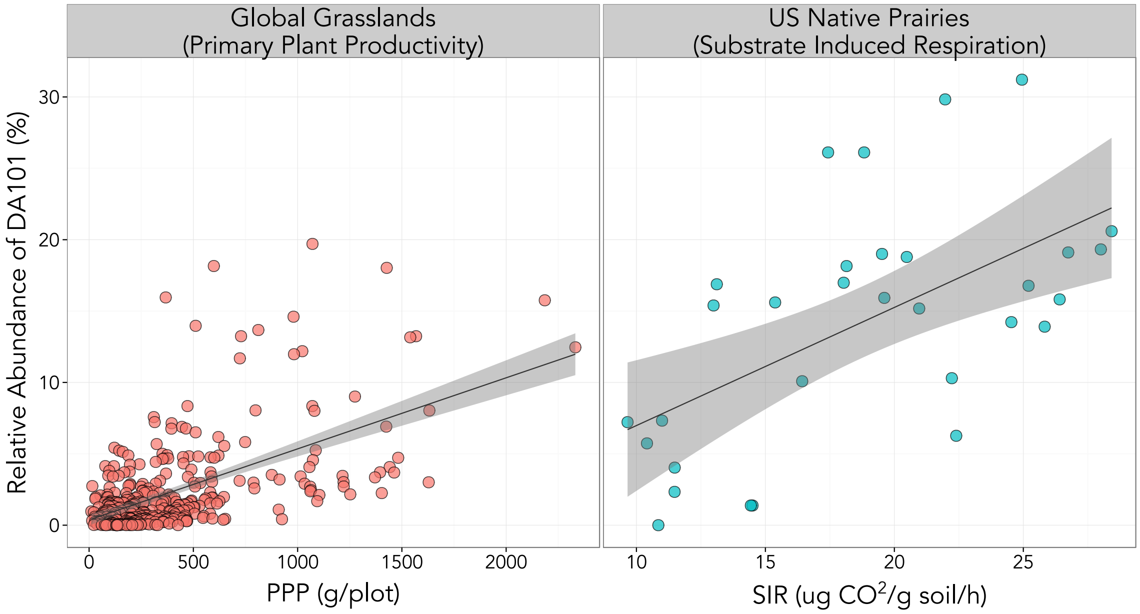
Acknowledgements: Funding to support this work was provided by grants from the National Science Foundation to N.F. (DEB0953331, EAR1331828), a Visiting Postdoctoral Fellowship award to P.C. from the Cooperative Institute for Research in Environmental Sciences, and an Alfred P Sloan Foundation grant to J.G. We also acknowledge infrastructural support provided by the University of Chicago Research Computing Center and University of Colorado Next Generation Sequencing Facility.



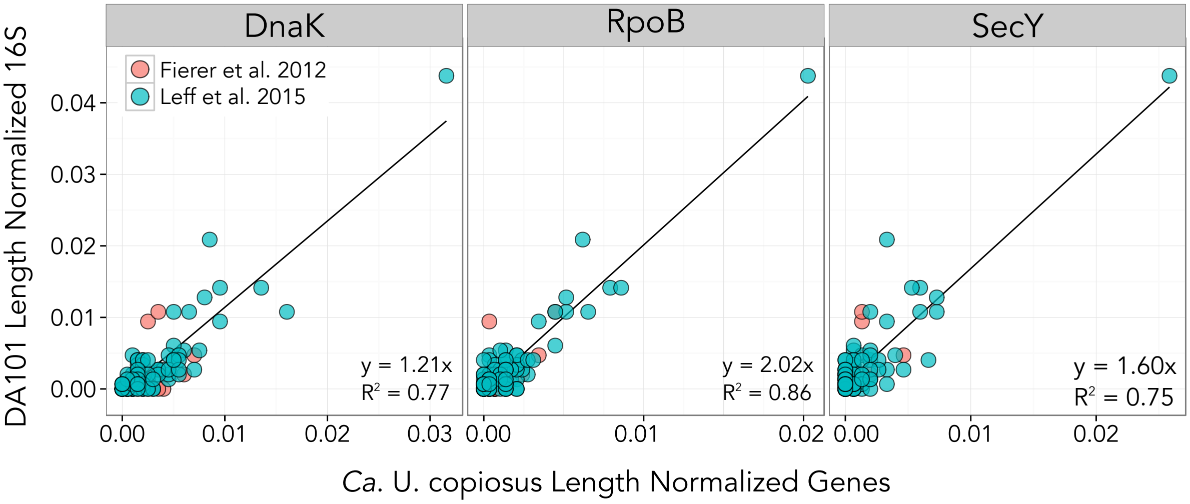
Supplemental Figure S3.1: DA101 rank is similar in amplicon and metagenomic data. The top 5 phylotypes from two matched amplicon and metagenomic datasets (Global Grasslands = Leff et al. 2015, Cross-Biome Soils = Fierer et al. 2012) are shown in order of decreasing median rank. Each point represents one sample within the corresponding dataset (Not all samples in the global grasslands dataset had metagenomic sequencing). DA101's position is highlighted with blue while all other phylotypes are grey.



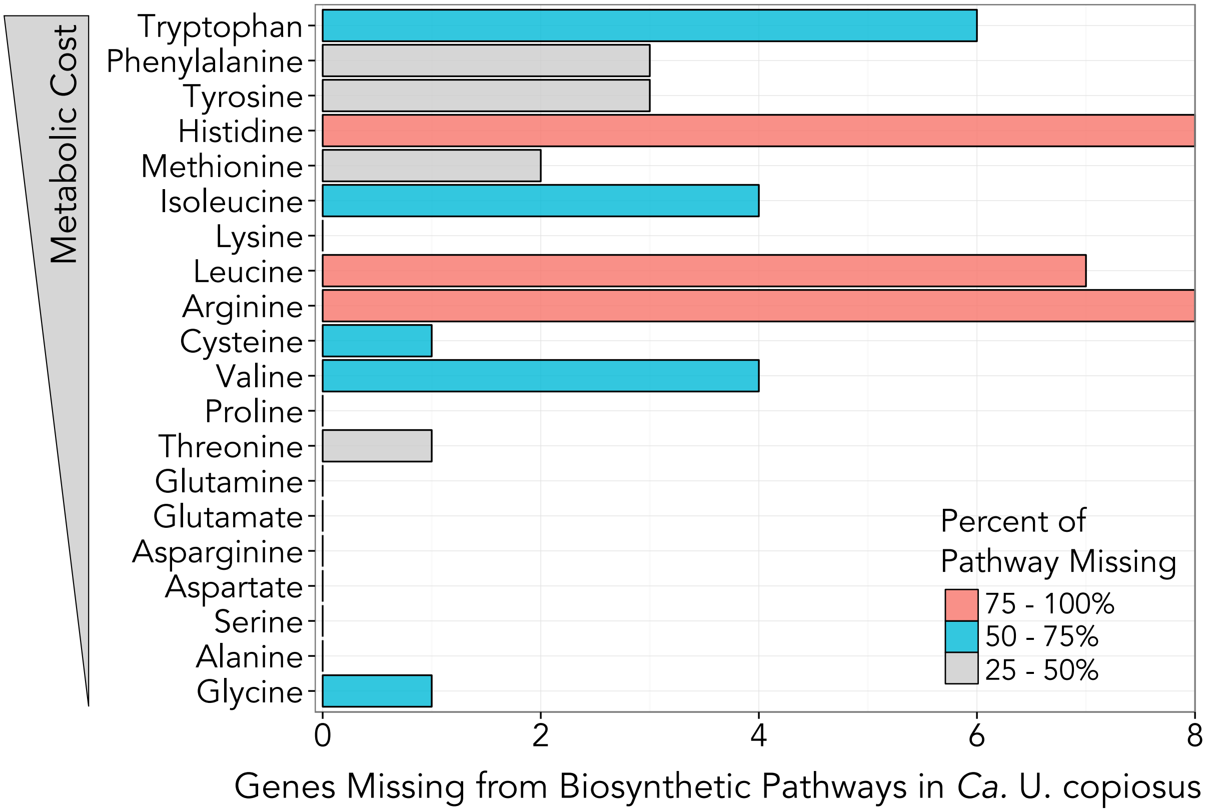
Supplemental Figure S3.2: Phylotype DA101 is more abundant in grasslands than forests. ($p<0.0001$, n=64, Mann-Whitney test) Data is from Crowther et al. 2014.



Supplemental Figure S3.3: The abundance of the DA101 amplicon correlates with measures of microbe and plant biomass. (Primary plant productivity p<0.0001 rho=0.47 n=366, and substrate induced respiration p<0.001 rho=0.57 n=31, Spearman correlations). The shaded region represents the 95% confidence interval of the trend line. Global Grasslands = Leff et al. 2015 and US Native Prairies = Fierer et al. 2013.



Supplemental Figure S3.4: The abundances of housekeeping gene sequences from the *Ca.* U. copiosus genome and DA101 16S rRNA gene sequences are well correlated across soil metagenomes ($P < 0.0001$, $\rho > 0.87$, n=102, Pearson correlation). We extracted matches to three *Ca.* U. copiosus housekeeping genes using blastN and compared the length normalized abundance of these fragments to the length-normalized abundance of DA101 16S rRNA gene sequences extracted using Metaxa2. We counted genes as a match to *Ca.* U. copiosus if the percent identity was greater than 85% and *Ca.* U. copiosus was the best hit. Our blastN database included the corresponding housekeeping genes from all named verrucomicrobial genomes in IMG. We chose 85% identity as our cutoff for several reasons: i) Protein coding genes are inherently more variable than rRNA genes; ii) the intraspecies percent identity variation for these genes has been reported to be as low as 87.7%; iii) there are no other representatives of this genus with a sequenced genome to permit direct comparisons.



Supplemental Figure S3.5: Pathways to synthesize several expensive amino acids are underrepresented in the *Ca.* U. copiosus genome. 34 unique genes are currently missing from the *Ca.* U. copiosus genome that would enable synthesis of all 20 amino acids. The cost of each of amino acid was estimated in E. coli by number of high-energy phosphate bonds hydrolyzed (Akashi and Gojobori, 2002). The number of genes missing in each pathway was calculated from KEGG metabolic pathways.

Supplemental Table S3.1: EMIRGE Samples

Sample Name	Dataset	Location	Description
NTP21	Fierer et al. 2013	Hayden, IA	Native prairie
NTP28	Fierer et al. 2013	Glynn Prairie, MN	Native prairie
NN1182	Leff et al. 2015	Val Mustair, Switzerland	Alpine grassland
NN772	Leff et al. 2015	Msunduzi Municipality, South Africa	Mesic grassland
TM25	New data set	Table Mountain, CO	Alluvial terrace
GG14	New data set	Gordon Gulch, CO	Meadow

Supplemental Table S3.2: Genome characteristics of heterotrophic soil Verrucomicrobia

Genome name	Lifestyle	Est. genome size (Mbp)	rRNA Copy #
Ca. Udaeobacter copiosus	Heterotroph	2.81	Likely 1
Chthoniobacter flavus Ellin428	Heterotroph	8.07	1
Opitutus terrae PB90-1	Heterotroph	5.96	1
Pedosphaera parvula Ellin514	Heterotroph	7.85	1

CHAPTER IV APPENDIX
UNLINKED RRNA GENES ARE WIDESPREAD AMONG ENVIRONMENTAL BACTERIA AND ARCHAEA

Materials and Methods

Analyses of complete genomes

All bacterial and archaeal genomes in the RefSeq genome database (O'Leary et al., 2016) classified with the assembly level "Complete Genome" were downloaded from NCBI in January 2019. We used gene ranges associated with each open reading frame (ORF) to pair the 16S and 23S rRNA genes that were closest to each other in each genome. We removed genomes that had an unequal number of 16S and 23S rRNA genes and those that had 16S or 23S on separate chromosomes or genome contigs (some genomes had up to 5 contigs). We also separated genomes that had a 16S rRNA gene within 1500bp of the end of the genome or a 23S rRNA genome within 1500bp of the beginning of the genome and classified these genomes independently. We classified the remaining rRNA pairs as either linked or unlinked if there was more than 1500bp between the end of the 16S and beginning of the 23S rRNA genes. Genomes were classified as 'unlinked', 'linked', or 'mixed' depending on the status of their rRNA operons with 'mixed' genomes having multiple rRNA copies with a combination of both linked and unlinked rRNA genes. All analyses were done in R version 3.5.1 (R Team, 2018).

Long-read shotgun metagenomic analyses

To investigate the prevalence of unlinked rRNA operon among those bacteria and archaea found in environmental samples (including many taxa for which genomes are not yet available), we analyzed long-read shotgun metagenomic datasets generated from soil, sediment, activated sludge, anaerobic digesters, and human gut samples. These metagenomic datasets were generated using either the Oxford Nanopore MinION (6 samples) or Illumina synthetic long-read sequencing technology (also known as Moleculo, first described here (Kuleshov et al., 2014), 9 samples). The Moleculo sequences originated from four separate studies: human gut (Kuleshov et al., 2016), prairie soil (White et al., 2016), sediment samples (Sharon et al., 2015), and grassland soils (MG-RAST project mgp14596, (Flynn et al., 2017). The MinION sequences originated from three separate unpublished studies featuring anaerobic digesters, activated sludge, and lawn soil. (Mads Albertsen & Arwyn Edwards). Drawing from these datasets, altogether we compiled 16,346,111 nanopore sequences and 890,542 moleculo (also known as Illumina synthetic long read) sequences. Altogether these 15 samples spanned multiple environments, from soil to sediment to anaerobic digesters to the human gut.

The first 250bp of each Nanopore sequence was removed because low quality. We performed no other quality filtering, as some samples did not include information on sequence quality (fasta format). We relied on our downstream filtering steps to remove sequences of poor quality. Metaxa2 was run on all sequences with default settings to search for SSU (16S rRNA) and LSU (23S rRNA) gene fragments. Taxonomy was assigned to the partial rRNA sequences using the RDP classifier (Wang et al., 2007) and the SILVA 123 SSU and LSU databases (Quast et al., 2012). Details on the number of reads per sample, read lengths, and the samples analyzed are available in Supplemental Table S4.1.

We next used a number of criteria to filter the reads included in downstream analyses and to identify taxa with unlinked rRNA genes. We only included those reads in our final dataset that: 1) included at least 2 domains of the 16S or 23S rRNA genes, 2) included either the last domain of the 16S rRNA gene or the first domain of the 23S rRNA gene, 3) the length of the 16S rRNA gene was ≤1591bp or the 23S rRNA gene was ≤ 3179bp (these limits correspond to the 99.5% quantile lengths of 16S and 23S rRNA gene sequences in our complete genomes), and 4) could be classified taxonomically to at least the phylum level. Of the subset of reads that met these criteria (39-347 per moleculo sample, 171-5071 per nanopore sample, see Supplemental Table S4.1 for details), we classified reads as containing an unlinked rRNA operon if there was ≥1500bp between the 16S and 23S, OR if there was no 23S domain found 1500bp after the end of the 16S rRNA. For our final analyses, we removed reads that could not be classified as linked or unlinked rRNA genes (for instance a sequence with only 300bp after the 3' end of the 16S rRNA gene) and included only reads that contained a 16S rRNA gene to avoid potentially double counting organisms with unlinked 16S and 23S rRNA genes. All analyses were done in R version 3.5.1 (Team, 2018).

Phylogenetic tree of unlinked rRNA genes

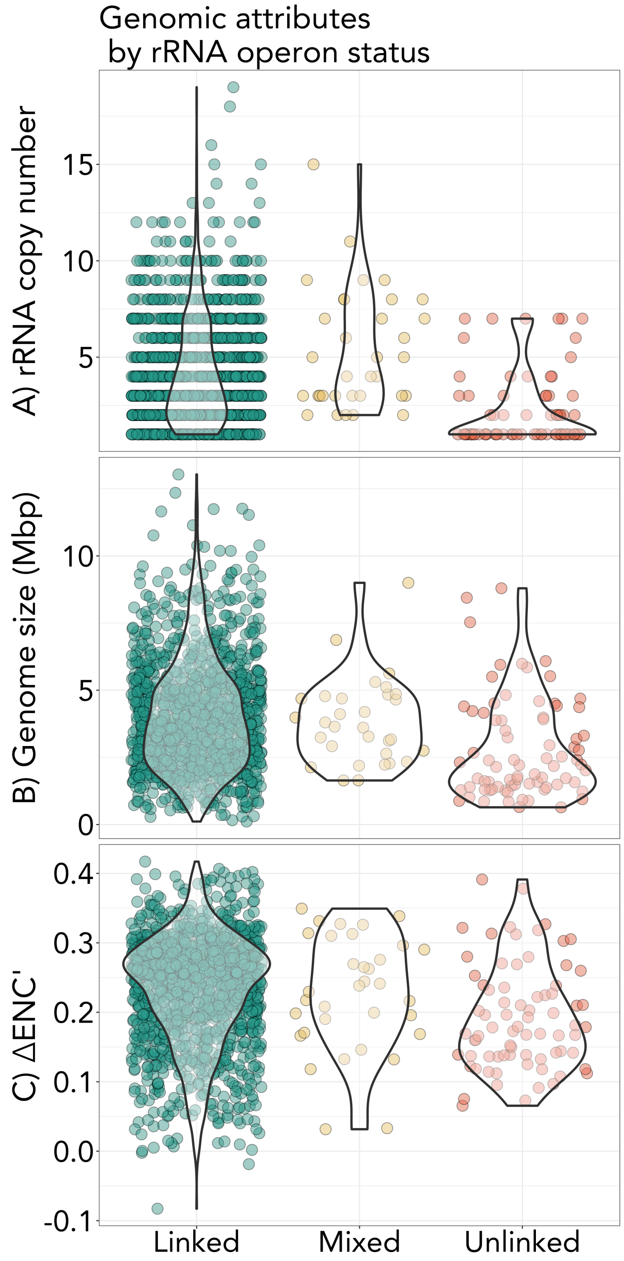
The phylogenetic tree (Figure 4.4) was created from full-length 16S rRNA sequences by combining both the NCBI complete genomes and the long-read shotgun metagenomic datasets. For the NCBI genome sequences, we selected one 16S rRNA gene sequence per unique species. For the long-read datasets, we first matched the partial 16S rRNA genes recovered by metaxa2 (Bengtsson-Palme et al., 2015) to full-length 16S rRNA gene sequences in the SILVA 132 SSU database (Quast et al., 2012) using the usearch10 (Edgar, 2010b) command usearch_global. Those full-length SILVA 16S rRNA genes sequences that matched to at least 95% percent identity were added to the complete genome sequences to represent the long-read metagenomic dataset. We used 95% percent identity as our cutoff as we found unlinked rRNA gene status to generally be conserved within genera. The NCBI and SILVA sequences were then aligned with PyNAST (Caporaso et al., 2010a) with a phylogenetic tree constructed using fasttree (Price et al., 2009), and plotted with ITOL (Letunic and Bork, 2016).

Genomic attributes associated with unlinked rRNA genes

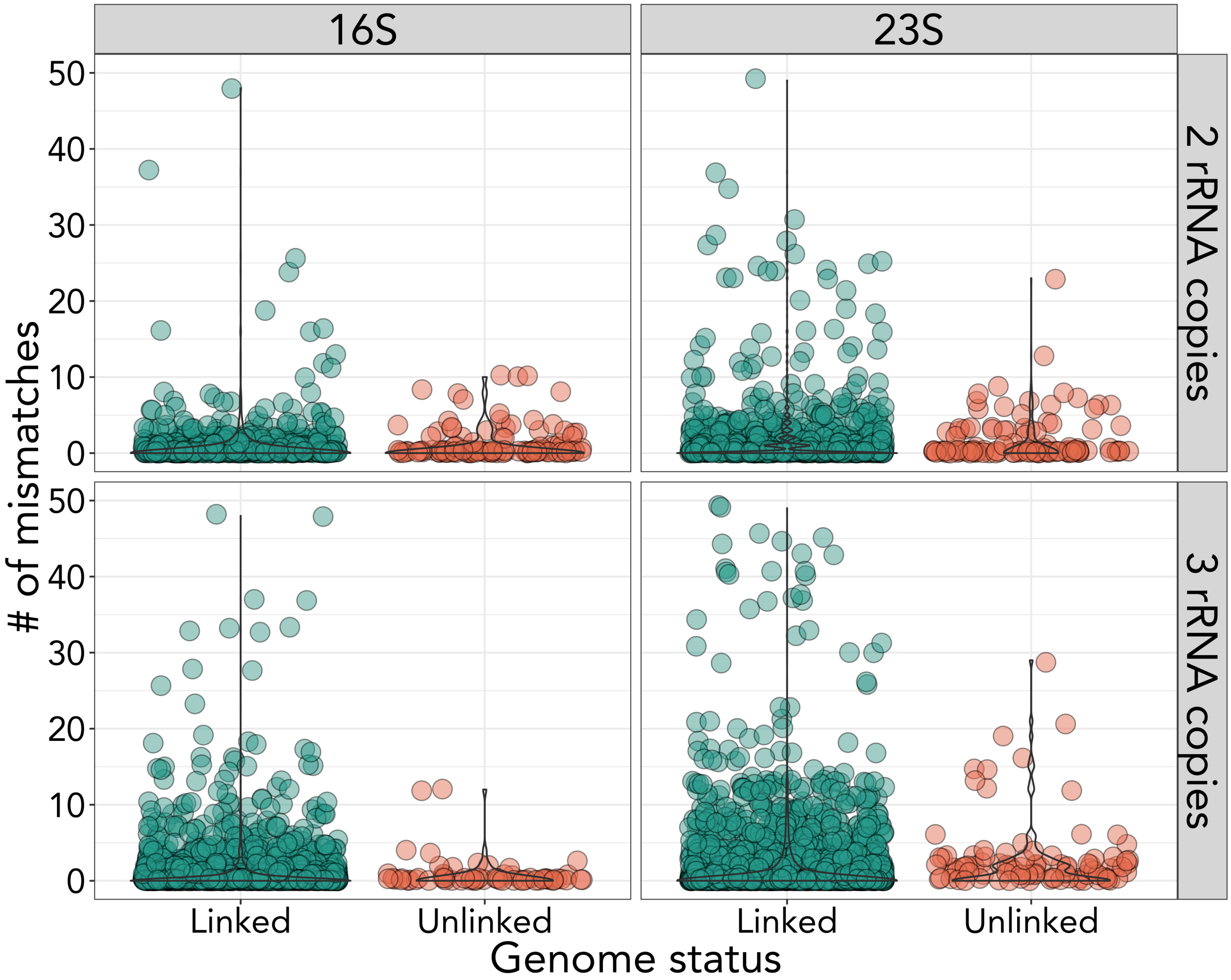
All tests for genomic attributes were done with a subset of our complete genome dataset - we reduced the dataset to include only one representative genome per unique species and operon status. For example, if a species had 24 genomes with linked rRNA genes and 3 genomes with unlinked rRNA genes, we retained two genomes total, one linked and one unlinked. To determine if taxa with unlinked rRNA genes have a lower predicted growth rate, we calculated the codon usage proxy ΔENC' (Novembre, 2002), which has been shown to provide an estimate of minimum generation times (Vieira-Silva and Rocha, 2009). We calculated ΔENC' with the program ENCprime (Novembre, 2002) with default options, on both the concatenated ORF sequences and concatenated ribosomal protein sequences for each genome following Vieira-Silva and Rocha, 2009. To check intragenomic rRNA sequence divergence, we used blastn with default settings to compare each pair of unique rRNA sequences in each of the genomes in our dataset. To determine if RNaseIII was present in each genome, we used hmmer (Eddy, 2011) to check for three RNaseIII pfams (bacterial PF00636, PF14622, and archaeal PF11469) in the translated protein files of each genome. We used the GA gathering cutoffs profile associated with each of these pfams to set all thresholding (--cut_ga).

Acknowledgements

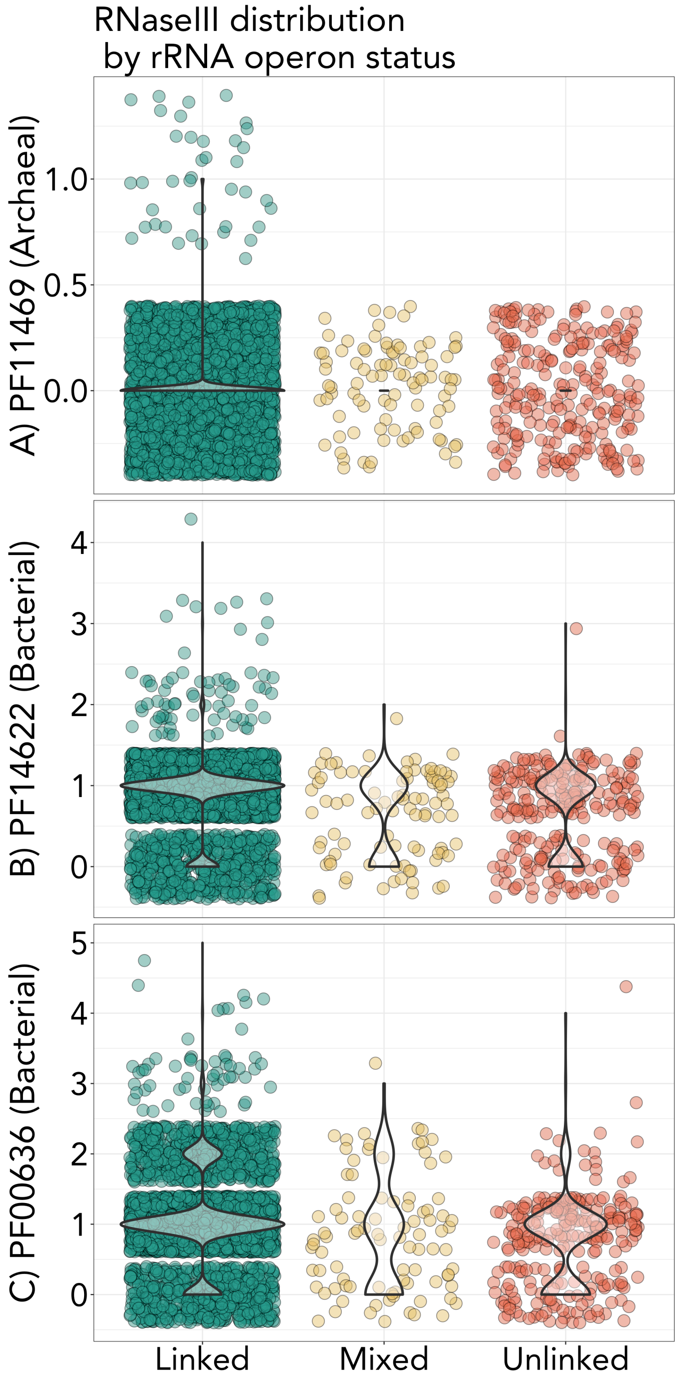
This research was supported in part by the Chateaubriand Fellowship of the Office for Science & Technology of the Embassy of France in the United States.



Supplemental Figure S4.1: Genomic attributes of NCBI complete genomes based on their rRNA operon status. Linked complete genomes feature exclusively linked rRNA genes, unlinked exclusively unlinked rRNA genes, and mixed have at least one linked and one unlinked rRNA operon. **A**) On average, genomes with exclusively unlinked rRNA genes had fewer rRNA copies (χ^2 p<0.001, means of groups: 4.2 linked, 5.5 mixed, 2.6 unlinked). **B**) Genomes with unlinked rRNA genes have smaller genomes on average, but this is not a significant difference (Supplemental Figure S1B, χ^2 p=0.87, means of groups: 4.1Mbp linked, 4.0 Mbp mixed, 2.8 Mbp unlinked). C) Genomes with exclusively unlinked rRNA genes are predicted to have a lower average generation time (χ^2 p<0.001, means of groups: 0.23 linked, 0.23 mixed, 0.19 unlinked). We calculated these statistics using a subset of our complete genomes with only one genome per unique species and operon status.



Supplemental Figure S4.2: rRNA divergence varies significantly between genomes based on their rRNA operon status. Sequence divergence among intragenomic 16S and 23S rRNA was significantly greater in genomes with unlinked rRNA (among genomes with 2-3 rRNA copies, χ^2 $p<0.05$ for 16S and 23S rRNA)



Supplemental Figure S4.3: Genomes with unlinked rRNA genes are less likely to encode bacterial RNaseIII. We found that there were significantly fewer RNaseIII genes in genomes with unlinked rRNA operons (PF00636: χ^2 p<0.001, means of groups: 1.0 linked, 0.84 mixed, 0.74 unlinked; PF14622: χ^2 p<0.001, means of groups: 0.86 linked, 0.63 mixed, 0.65 unlinked). We also checked this relationship for archaeal RNaseIII, but found no significant association (PF11469: χ^2 p=0.153). We calculated these statistics using a subset of our complete genomes with only one genome per unique species and operon status.

Supplemental Table S4.1: Sequence statistics

Sample (sequence type)	Sequences	Median length	Total 23S	23S passing filter (%)	Total 16S	16S passing filter (%)
Lawn soil (n)	1751625	2706	2767	19.05	2085	28.82
Anaerobic digester 2 (n)	1462320	6038	6196	22.27	5276	30.34
Anaerobic digester 1 (n)	3362711	2770	13577	15.50	10910	23.24
Anaerobic digester 3 (n)	6194277	5393	18389	21.75	15668	32.37
Activated sludge (n)	1366686	7875	4152	6.17	3412	5.01
Sediment 4 (n)	2208492	5787	3550	21.63	2869	22.93
Grassland soil 4 (m)	50850	3305	97	47.42	63	69.84
Grassland soil 1 (m)	67177	9618	136	63.24	132	76.52
Grassland soil 2 (m)	115256	7022	256	45.70	223	65.92
Grassland soil 3 (m)	34170	5861	73	54.79	72	54.17
Sediment 1 (m)	9282	7863	232	53.02	196	72.96
Sediment 2 (m)	13190	7317	274	48.91	253	62.06
Sediment 3 (m)	9282	7859	258	54.65	187	74.87
Human gut (m)	65354	7808	692	50.14	534	60.49
Grassland soil 5 (m)	123687	7197	248	61.29	213	74.65