DOE Final Technical Report

1. DOE Award Number: DE-SC0010566

2. Recipient: University of California, Berkeley

3. Project Title: Multi-'omic' analyses of the dynamics, mechanisms, and pathways for carbon turnover in grassland soil under two climate regimes

4. Principal Investigator: Prof. Jillian F. Banfield, Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, CA, USA

5. Date of Report: March 28, 2019

6. Period Covered by Report: 2013 – 2019

7. Research Accomplishments

Soil microbial activity drives the carbon and nitrogen cycles and is an important determinant of atmospheric trace gas turnover, yet most soils are dominated by microorganisms with unknown metabolic capacities. This knowledge gap precludes meaningful predictions of relationships between microorganism types and their biogeochemical functions. Specifically, as climate change physically alters terrestrial ecosystems, the impact on microbial consortia and potential consequences for release of carbon and other nutrients from soils are difficult to predict. In this project, we investigated the identities, activity, and functional potential of microorganisms residing in a well characterized grassland ecosystem (the Angelo Coast Range Reserve) undergoing a controlled 14-year rainfall extension climate change experiment **(Fig. 1)**. The research was achieved using a combination of "omics"-based technologies, including genome-resolved metagenomics, transcriptomics, and proteomics. The approach allowed us to capture multiple facets of community ecology and activity *in situ*. Supported by this award, significant progress was made in addressing the grand challenge of achieving genome-resolution for very complex soil microbial communities. In addition to improvements in the methodology, the research required new approaches to define the 3-dimentional distributions of microorganisms and their metabolic traits in soils and to decipher the impact of climate. Underpinning such analyses was the need for sufficient statistical power to test for geographic, soil depth and rainfall controls on community composition and metabolic properties. This required that we scale up throughput substantially. We overcame significant technical barriers to achieve complex microbial community resolution.

The work uncovered previously largely overlooked phenomena, including

(a) the importance of methylotrophy independent of methane oxidation
(b) the importance of lanthanide dependent enzymes for methyl group oxidation
(c) the prevalence of CO oxidation
(d) the prominence of secondary metabolism in soil
(e) the importance of organisms not previously linked to secondary metabolite production

(f) ammonia oxidation by unexpected archaeal groups
(g) overall prevalence of newly described candidate phyla in soil
(h) depth stratifies carbon and nitrogen turnover capacity, with small molecule processing more prevalent near the surface and inorganic nitrogen turnover more prevalent at depth
(i) climate change-related rainfall extension directly impacts genome and metabolic functional distribution with decreases in inorganic nitrogen turnover at deeper depths and a corresponding increase in complex carbohydrate utilization capacity

These findings are summarized in three major publications and two papers currently in preparation:

A. **Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone.** Butterfield CN, Li Z, Andeer PF, Spaulding S, Thomas BC, Singh A, et al. (2016) ***PeerJ***, *4*(3), e2687–28. http://doi.org/10.7717/peerj.2687

In this work, we overcame many of the initial technical barriers for performing genome-resolved metagenomics, shotgun proteomics, and metabolomics in a complex soil system. From ten soil samples, we reconstructed 198 genomes that span all major phyla previously detected at this site via marker gene studies (Cruz et al. 2009). We also quantified 6,835 proteins and 175 metabolites and showed that after a natural rainfall event the concentrations of many sugars and amino acids approach zero at the base of the soil profile. Most prior research on carbon cycling in soil has focused on microbial degradation of complex soil organic macromolecules, likely derived in part from plant biomass. However, the metabolomics, proteomics, and metagenomic analysis in this study all suggest that small organic compounds, particularly C1 compounds, are likely a significant and widely utilized energy source for soil microorganisms. Quantitative and functional profiling of resolved genomes indicated that bacteria from the poorly sampled phyla, Gemmatimonadetes and Rokubacteria, were relatively abundant at the site, and encode a significant capacity for C1 carbon metabolism via methylotrophy. Our proteomics data also reflected these findings, as proteins for methanol oxidation were among the most abundant proteins in the proteome. Interestingly, the methanol dehydrogenases identified by both proteomic and genomic analysis were by and large a subtype known as XoxF, which have been identified relatively recently and are known to bind lanthanide ions as cofactors. The widespread distribution of XoxF type methanol dehydrogenases at this site suggest that, at least in some soil systems, lanthanides may be important biological co-factors for carbon turnover activity. This work also established that previously undescribed archaea from the phyla Bathyarchaeota and Thermoplasmatota, are abundant in deeper soil horizons and are inferred to contribute appreciably to aromatic amino acid degradation. Overall, this genome-resolved multi-omic study revealed many populations of little known bacteria and archaea in sub-root zone soil microbial communities, and highlighted previously little considered metabolites and nutrient co-factors as having important roles in soil microbial ecosystems.

B. **Mediterranean grassland soil C-N compound turnover is mediated by genomically divergent organisms, depth stratified, and rainfall dependent.** Diamond S, Andeer PF, Li Z, Crits-Christoph A, Burstein D, Anantharaman K, Lane K, Thomas BC, Pan C, Northen TR, and Banfield JF. (2019) ***Nature Microbiology*** (Accepted).

In this work, we applied deep metagenomic sequencing and metaproteomic analyses to sub-root zone samples from the Angelo Coast Range reserve. Our specific goal was to find statistically supported differences in organism phylogenetic and functional distribution across soil depths and between un-treated and treated soil plots exposed to the 14-year rainfall extension climate change experiment. We resolved sixty metagenomic and twenty proteomic datasets, recovered 793 near-complete microbial genomes from 18 phyla, and identified 55,665 proteins with at least one uniquely mapped peptide across our samples. Our genomic coverage of organisms from a soil system was unpresented for the time, as we were able to recover genomes for >50% of detected microorganisms our soil, based on coverage (which is a measure of cells sampled). Overall our data revealed important carbon and nitrogen turnover functions in understudied microbial groups, showed a stark metabolic and phylogenetic stratification across soil depths, and supported climate change-based rainfall extension as a factor that can significantly alter the carbon and nitrogen turnover capacity of soil microbial communities.

Again, we found that Lanthanide cofactor bearing XoxF-type methanol dehydrogenases were highly prevalent, and the only methanol dehydrogenase class identified at our site. We additionally detected a large number of credible type-I coxL carbon monoxide dehydrogenases in both genomic and proteomic data, supporting CO as an important C1 energy source in these soils. However, we also noted a high number of phylogenetically related coxL-like sequences that likely function as small molecule dehydrogenases on other substrates. We posit these coxL-like enzymes may play uncharacterized roles in plant exudate processing and turnover in the studied soils, and have provided their sequence and associated genome data to the scientific community for further study. We also observed that a large fraction of the enzymes typically involved in complex carbohydrate degradation were carbohydrate esterases, which cleave methyl and acetyl groups from complex carbohydrate polymrs. As methyl and acetyl groups are common additions to many polymers, the widespread prevalence of carbohydrate esterases may represent a strategy where readily available C1 and C2 carbon is accessed with minimal energetic investment. This observation may explain, in part, why low molecular weight carbon molecules are important currencies in this ecosystem.

Enzymes involved in complex carbon metabolism, C1, and small molecule turnover were statistically enriched in microorganisms closer to the surface, suggesting that metabolic strategies at shallow depths are structured around plant-derived exudates and complex carbon. These data support previous observations that soil organic matter has significantly shorter residence time closer to the soil surface. In contrast, most inorganic nitrogen transformation functions are more prevalent or exclusively found in microorganisms enriched at greater depth. Thus, N_2O discharged to the atmosphere from Mediterranean grasslands may originate from deeper soil strata.

Under the treatment involving extended Spring rainfall, the relative decrease of microorganisms at deeper depths performing ammonia liberation and oxidation suggests a mechanism by which climate change could limit nitrogen cycling and N_2O release. Simultaneously, increased complex carbohydrate degradation capacity at deeper soil depth could counter this climate change impact by increasing CO_2 release from normally recalcitrant soil organic matter. However, the kinetics of CO_2 and N_2O release in response to rainfall changes, and the generality of these findings to other soils, remain uncertain. What is certain is that climate change can have a direct impact on the relative abundance and metabolic capacities of microorganisms in soil ecosystems, with potentially important impacts for trace gas release.

C. **Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis**. C. Crits-Christoph A, Diamond S, Butterfield CN, Thomas BC, and Banfield JF (2018) ***Nature***, 558: 440-444. doi: https://doi.org/10.1038/s41586-018-0207-y

In this study, using genomes reconstructed during study B, we identified microorganisms from previously understudied phyla that encode diverse polyketide (PKS) and non-ribosomal peptide (NRPS) biosynthetic gene clusters. The biosynthetic loci are encoded by newly identified members of the Acidobacteria, Verrucomicobia, Gemmatimonadetes, and Rokubacteria, which are groups known to be highly abundant in soils but not previously linked to secondary metabolite production with confidence. We also demonstrate that these PKS and NRPS biosynthetic clusters are divergent from known types, and use metatranscriptomic from microcosms constructed with Angelo Coast Range Reserve soils to show that many of these gene clusters are co-expressed and show differential transcriptional responses to soil nutrient inputs. To our knowledge this is the first application of genome-resolved metatranscriptomics applied to a soil system. Also by using whole genomes in this study, we were able to link the novel biosynthetic clusters to the functional capacities and ecological context of their encoding genomes. Specifically, we were able to construct transcriptional co-expression networks for a number of the organisms investigated that demonstrated some biosynthetic gene clusters were co-expressed with two-component systems, transcriptional activators, antimicrobial resistance, and iron regulatory genes.

Across the 376 genomes surveyed we identified a total of 1,159 biosynthetic gene clusters on contigs at least 10 kbp in length. No protein in any biosynthetic cluster identified shared more than 79.7% amino acid identity across ≥50% of the full protein lengths with reference proteins from the Minimum Information about a Biosynthetic Gene (MIBiG) repository. Fifty-nine per cent of predicted proteins had no ≥50%-length homologue in MIBiG, and those that did shared an average of only about 39% amino acid identity to the best hit of any MIBiG protein. Using the same thresholds for gene homologues, we found that 220 clusters did not share more than 50% of the genes of any previously described cluster. While we were not able to assign specific functions to most of the clusters we identified, we found strong evidence that many of the NRPS and PKS clusters (242 total) may be involved in producing antimicrobial compounds as 153 proteins found within 84 (out of 240) of our NRPS and PKS clusters encoded transporters known to function in antimicrobial compound resistance.

Two near-complete draft Acidobacterial genomes also encoded unusually large repertoires of NRPS and PKS genes. The *Candidatus Eelbacter* and *Candidatus Angelobacter* genomes encoded 17 and 16 biosynthetic loci respectively. The discovery of these two microorganisms establishes that bacterial specialization in secondary metabolite biosynthesis is not limited to known clades in the Actinomycetales, Proteobacteria, Cyanobacteria, Bacilli and the recently discovered Entotheonella. As our organisms originated from a microbially complex environment, their existence is in agreement with proposed ecological and evolutionary forces, such as cooperative and competitive lifestyles, that select for organisms with high secondary biosynthetic capacity.

By analyzing metatranscriptomics data from 120 soil microcosm samples derived from our field site that were subject to amendment with glucose, methanol or water over 24 h, we probed the strong biological responses that occur in soils following water addition and nutrient release after a long dry period. Overall we detected expression from 133 out of 242 NRPS and PKS clusters in our full dataset, and found that co-expression of genes within all biosynthetic clusters was significantly more common than co-expression of within cluster genes and other genes in an

organism's respective genome (Wilcoxon rank-sum test, $P < 0.001$). In looking at genomes with temporally significant gene expression patterns across our microcosim experiment we found distinct expression clusters enriched in secondary metabolic genes. These clusters also contained genes involved in two-component systems, efflux and transcriptional regulators, and were almost completely devoid of genes for the core processes of transcription, translation and energy metabolism. Specifically in *Candidatus Angelobacter*, genes from five NRPS/PKS clusters were co-expressed together in a module with a variety of genes involved in environmental sensing and response, including homologues of the iron siderophore uptake receptor TonB, homologues of the macrolide export trans-porter MacB, two putative antimicrobial resistance genes, and an operon for a type VI secretion system.

Overall, we uncovered extensive evidence for secondary metabolite synthesis in a large collection of bacterial genomes from four phyla of soil bacteria that have not previously been genomically linked to this capacity. Although specific functional assignment by homology was difficult for these novel gene clusters, the transcriptional associations between specific NRPS and PKS gene clusters, regulators of iron metabolism, and putative antimicrobial resistance mechanisms suggest that at least some clusters be involved in competition for iron resources and antibiotic production.

Work in preparation: Due to the high depth and breadth of our "omics" datasets, we have continued to extract valuable biological information from these samples while they exist in the public domain. As a result, two pending, but currently unpublished, studies will be released using the data generated by this funding. These studies address the identification of a highly novel archaeal lineage that may function in nitrogen turnover, (E) and the prevalence and impact of genomic strain variation across the Angelo soil sampling site (PhD research of Alex Crits-Christoph.

D. **A Divergent Particulate Monooxygenase Identified in a Novel Archaeal Lineage**. Spencer Diamond, Adi Lavy, Alex Crits-Christoph, Allison Sharrar, Evan Star, Paula M Carnevali and Jillian F. Banfield. *In preparation* for submission very soon.

We identified a highly novel archaeal lineage encoding particulate methane/ammonia monooxygenase genes. Currently the only known archaeal lineages to encode this function are within the Thaumarchaeota superphylum. The novel Archaea we identified, *Candidatus* Angelarchaeales, fall within the Euryarchaeota superphylum, and their identification significantly expands the range of Archaea with the capacity to perform this oxidative function. The particulate monooxygenases in *Candidatus* Angelarchaeales have been confirmed both at the sequence and phylogenetic level, and all observed instances fall within an expected 4 gene pmoABCX operon.

As *C.* Angelarchaeales were a relatively abundant lineage in Angelo soils, and because we also detect related organisms from this lineage at other soil and subsurface sites, we propose this may be a widely prevalent and overlooked group contributing to ammonia or methane turnover. As most environmental studies of ammonia and methane oxidation genes use primer-based surveys, we were not surprised these organisms have been overlooked as their monooxygenase genes are highly divergent from known sequences. Also, while we have not biochemically characterized the exact substrate used by these organisms, we note that many isolated particulate methane/ammonia monooxygenases act on both methane and ammonia. However, we also will provide genomic metabolic context, such as the presence of nitrite oxioreductase genes in some

strains, that suggest the function of these novel monooxygenases in the turnover of ammonia. Overall this finding could have broad implications for understanding the turnover of nitrogen in soils and subsurface environments.

8. Participants

Graduate Students: Allison Sharrar (**30** %), Alex Crits-Christoph (**30** %)

Postdocs: Christina Butterfield (**100** %), David Burstein (**20** %), Spencer Diamond (**100** %)

Staff: Susan Spaulding (**20** %), Brian C Thomas (**20** %), Kathrine Lane (**10** %)

David Burstein is now an Assistant Professor at University of Tel Aviv in Israel, Christina Butterfield became a scientist at "2nd Genome" and is now a scientist at the "Metagenomi" startup founded primarily by Brian Thomas. Sue Spaulding went to nursing school and Kate Lane accepted admission to the UC Davis microbiology PhD program. Allison Sharrar is now a staff member at UC Berkeley, Alex Crits-Christoph is current a UC Berkeley PhD student and Spencer Diamond is a postdoc working on the DOE-funded mCAFEs project led by LBNL.

9. Unexpended Funds: The grant is fully expended

Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone

Cristina N. Butterfield[1], Zhou Li[2], Peter F. Andeer[3], Susan Spaulding[1], Brian C. Thomas[1], Andrea Singh[1], Robert L. Hettich[2], Kenwyn B. Suttle[4], Alexander J. Probst[1], Susannah G. Tringe[5], Trent Northen[3], Chongle Pan[2] and Jillian F. Banfield[1,3]

[1] Department of Earth and Planetary Sciences, University of California, Berkeley, CA, United States
[2] Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, Unites States
[3] Lawrence Berkeley National Laboratory, Berkeley, CA, United States
[4] Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, CA, United States
[5] DOE Joint Genome Institute, Walnut Creek, CA, United States

ABSTRACT

Annually, half of all plant-derived carbon is added to soil where it is microbially respired to CO_2. However, understanding of the microbiology of this process is limited because most culture-independent methods cannot link metabolic processes to the organisms present, and this link to causative agents is necessary to predict the results of perturbations on the system. We collected soil samples at two sub-root depths (10–20 cm and 30–40 cm) before and after a rainfall-driven nutrient perturbation event in a Northern California grassland that experiences a Mediterranean climate. From ten samples, we reconstructed 198 metagenome-assembled genomes that represent all major phylotypes. We also quantified 6,835 proteins and 175 metabolites and showed that after the rain event the concentrations of many sugars and amino acids approach zero at the base of the soil profile. Unexpectedly, the genomes of novel members of the Gemmatimonadetes and Candidate Phylum Rokubacteria phyla encode pathways for methylotrophy. We infer that these abundant organisms contribute substantially to carbon turnover in the soil, given that methylotrophy proteins were among the most abundant proteins in the proteome. Previously undescribed Bathyarchaeota and Thermoplasmatales archaea are abundant in deeper soil horizons and are inferred to contribute appreciably to aromatic amino acid degradation. Many of the other bacteria appear to breakdown other components of plant biomass, as evidenced by the prevalence of various sugar and amino acid transporters and corresponding hydrolyzing machinery in the proteome. Overall, our work provides organism-resolved insight into the spatial distribution of bacteria and archaea whose activities combine to degrade plant-derived organics, limiting the transport of methanol, amino acids and sugars into underlying weathered rock. The new insights into the soil carbon cycle during an intense period of carbon turnover, including biogeochemical roles to previously little known soil microbes, were made possible via the combination of metagenomics, proteomics, and metabolomics.

Subjects Bioinformatics, Environmental Sciences, Genomics, Microbiology, Soil Science

Submitted 10 August 2016
Accepted 14 October 2016
Published 8 November 2016

Corresponding author
Jillian F. Banfield,
jbanfield@berkeley.edu

Academic editor
A. Murat Eren

Additional Information and Declarations can be found on page 19

DOI **10.7717/peerj.2687**

How to cite this article Butterfield et al. (2016), Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone. **PeerJ 4:e2687; DOI 10.7717/peerj.2687**

Keywords Genome-resolved metagenomics, Methanol dehydrogenase, Soil bacteria, Soil archaea, Proteomics, Metabolomics

INTRODUCTION

The terrestrial carbon reservoir is primarily distributed among grasslands (∼34%), forests (∼39%), and cultivated farms (∼17%) (*White, Murray & Rohweder, 2000*). While in forests the majority of fixed carbon is stored vegetation, most fixed carbon in grasslands is stored in soil. Thus, grassland soils are one of the most important reservoirs of terrestrial carbon on the planet. This observation has motivated many studies of processes that impact the fate of carbon compounds in grassland soil systems (*Blazewicz, Schwartz & Firestone, 2014*; *Kandeler et al., 2006*; *Mau et al., 2015*; *Reinsch et al., 2014*; *Schimel & Schaeffer, 2012*; *Verastegui et al., 2014*). Over a variety of time scales, organic detritus is either respired back to CO_2 or degraded into smaller molecules that are transported in solution to underlying zones (*Placella, Brodie & Firestone, 2012*). In Mediterranean climate soils, this process emits an amount of CO_2 equal to the annual output from other ecosystems immediately following the first Fall rain, when nutrients from senesced plants are driven downward. Carbon exported from the shallow soils provides nutrients for microorganisms in lower soil horizons and the deeper subsurface.

Understanding which microorganisms are present and the pathways by which they process carbon compounds is key to understanding the form and redistribution of soil organic matter. Soil ecologists and microbiologists have employed, and continue to employ, isolation, phospholipid-derived fatty acid (PLFA) analysis, fingerprinting with denaturing gradient gel electrophoresis (DGGE) (*Muyzer, De Waal & Uitterlinden, 1993*) and terminal fragment length polymorphism (T-RFLP) (*Osborn, Moore & Timmis, 2000*) analyses, and 16S rRNA amplification and sequencing (*Banning et al., 2011*; *He, Xu & Hughes, 2006*; *Henckel, Friedrich & Conrad, 1999*) to identify soil microbes. Bacteria reported in soils are typically affiliated with the Proteobacteria, Firmicutes, Acidobacteria, Actinobacteria, Verrucomicrobia, Gemmatimonadetes, and Bacteroidetes phyla (*Evans & Wallenstein, 2012*; *Fierer et al., 2012*; *Kuramae et al., 2012*). Given the evolutionary/adaptation pressures in different soil types (*McKissock, Gilkes & Walker, 2002*), nutrient availability (*Adair, Wratten & Lear, 2013*; *Goldfarb et al., 2011*; *Veresoglou et al., 2012*), temperature and moisture (*Aanderud et al., 2013*; *Peltoniemi et al., 2015*), pH (*Lauber et al., 2009*), variety of vegetation (*Herzberger, Duncan & Jackson, 2014*; *Piper et al., 2015*; *Prober et al., 2015*), and microenvironments within the soil, it is expected that there will be substantial genetic diversity within many of these reported phyla, giving rise to clusters of closely related organisms, as well as organisms from additional phyla that have thus-far eluded detection in soils.

Prior studies of microbial functions (i.e., ammonia oxidation, methylotrophy) in soil have used targeted approaches such as gene amplification (qPCR, pyrosequencing) (*Hofmann et al., 2016*; *Pester et al., 2012*; *Stacheter et al., 2013*), culturing of isolates and enrichments (*Beck et al., 2014*). More recently, metagenomic methods have been applied to soil samples with the objective of providing a cultivation- and primer-independent

view of microbial community composition and functional capacities (*Delmont et al., 2015*; *Hultman et al., 2015*; *Luo et al., 2014*; *White et al., 2016*). Genomes provide metabolic insight (including for organisms that have not been cultivated) and enable identification of pathways involved in biogeochemical processes, but they have rarely been reconstructed from soil (*Delmont et al., 2015*; *Hultman et al., 2015*; *Pell et al., 2012*). These recent studies have shown that resolution of the community is possible and that much of the community's metabolic potential centers around respiration, complex carbohydrate degradation and central metabolism.

Here we conducted a multi-omic investigation of the microbiology and microbial activity in the shallow sub-root and deeper regions of grassland soil that experiences a Mediterranean climate. Samples were collected during the major annual period of carbon turnover around the time of the first Fall rain event. The well-studied meadow (*Cruz-Martínez et al., 2009*; *Suttle, Thomsen & Power, 2007*) is located in the Angelo Coastal Reserve in Northern California (part of the Eel River Critical Zone Observatory). These grassland soils provide an ideal system for studying processes in the sub-root zone because roots are confined to a well-defined horizon. Prior work in this meadow documented that carbon is fixed by ~50 plant species (*Suttle, Thomsen & Power, 2007*) during the wet spring season. The carbon compounds accumulate in the upper soil horizon after plants die late in summer (*Aerts, Bakker & De Caluwe, 1992*; *Berendse, 1994*; *Wedin & Tilman, 1990*). Thus, portions of these carbon compounds that accumulate in the upper 10 cm of the soil include both leaf litter and dead root material will percolate down as dissolved metabolites through the 40–50 cm deep soils, which are developed on weathered vermiculite-dominated argillite and sandstone.

Our sampling scheme was designed to probe microbial diversity and active carbon turnover in soil using a combined metagenomic, proteomic and metabolomic approach. An important motivation for recovery of genomes from the metagenomes is that protein sequences can be predicted in organism context and used in mass spectrometry studies to identify proteins that are highly abundant in microbial cells (*Brooks et al., 2015*; *Mosier et al., 2015*). This information, in combination with metabolite concentrations measured through the soil profile, enables identification of the organisms, pathways and spatial distribution of carbon turnover processes at the time of sample collection. We uncovered roles for bacteria and archaea from phylum lineages lacking isolated representatives and identify methylotrophy and archaeal heterotrophy as major carbon cycling processes in the sub-root zone. The study demonstrates that genome-resolved multi-omic approaches can be effectively used to interrogate microbiallymediated processes in one of the Earth's complex ecosystems.

MATERIALS & METHODS

Sampling and DNA extraction

We collected samples from the Angelo Coast Range Reserve (with permission under APP # 27790) meadow 39°44′21.4″N 123°37′51.0″W) on four days: before the rain, four and six days after one inch of rain fell, and two days after three inches of rain fell in September, 2013.

At two plots 10 m apart in the Northern end of the meadow, a soil pit for each sampling day was dug to 50 cm and depth was marked every 10 cm. Approximately 1 kg of soil was removed from each of two depths (10–20 cm and 30–40 cm) using sterilized stainless steel hand trowels. Each sample was homogenized briefly in a sterile bowl and divided into several sterile Whirl-Pak bags. One sample bag was placed on wet ice for transport to the lab for pH and moisture analyses. The remaining samples were immediately flash frozen in a mixture of dry ice and ethanol and then placed on dry ice for transport to the lab for long-term storage at −80 °C. For pH analysis, 2 g of fresh, field-wet soil were suspended in 10 mL 0.01 M $CaCl_2$ and shaken for one hour at 100 rpm. The suspension was then centrifuged for 5 min at 6,000 rpm at 4 °C, and the resulting supernatant was filtered through a #1 Whatman filter and analyzed with a pH probe. Gravimetric soil moisture was determined by weighing subsamples of sieved soil (2 mm sieve) before and after drying at 60 °C for at least 48 h. Soil particle size distribution was determined by measuring the relative density of the soil suspended in 5% sodium hexametaphosphate solution with a hydrometer over the course of settling (40 s–2 h). To measure extractable organic carbon content, 5 g soil samples were suspended in 25 mL of 0.5 M potassium sulfate and shaken for two hours at 150 rpm. The suspension was then filtered through a #1 Whatman filter and quantified on an OI Analytical 1010, Total Organic Carbon Analyzer. Soil characteristics are summarized in Fig. S1.

For each depth, DNA was extracted using MoBio Laboratories PowerMax Soil DNA Isolation kits from 10 g of soil from ten of a the much larger set of samples used for other analyses: (1) pre-rain in plot 1 from 10–20 cm, (2) pre-rain in plot 1 from 30–40 cm, (3) four days after the first rain in plot 1 from 10–20 cm, (4) four days after the first rain in plot 1 from 30–40 cm, (5) six days after the first rain in plot 1 from 10–20 cm, (6) six days after the first rain in plot 1 from 30–40 cm, (7) six days after the first rain in plot 2 from 10–20 cm, (8) two days after the second rain in plot 1 from 10–20 cm, (9) two days after the second rain in plot 1 from 30–40 cm, and (10) two days after the second rain in plot 2 from 10–20 cm. We optimized the protocol for our samples, to maximize DNA yield while minimizing shearing: each sample was only vortexed for 1 min, followed by a 30 min heat step at 65 °C, inverting every 10 min. Each sample was also extracted twice, and combined at the spin filter step. We performed two elution steps of 5 mL each, and precipitated the DNA using sodium acetate and glycogen, resuspending in 100 uL of 10 mM Tris buffer. This resulted in unsheared large fragment size DNA, with average yields of 2,351 ng/g soil (10–20 cm depth) and 1,277 ng/g soil (30–40 cm depth). Fragment size was checked on 0.5% agarose gels using a 23 kb genomic DNA ladder and DNA concentration was measured using a Qubit Fluorometric Quantitation device, dsDNA Broad Range Assay Kit.

DNA sequencing and reconstruction of genomes

DNA sequencing was conducted at the Joint Genome Institute, USA. 250 bp paired Illumina reads were processed with BBMap (https://sourceforge.net/projects/bbmap/). BBMap was run twice (1) to trim adapters bbduk.sh was used with parameters $k = 23$, mink = 11, hdist = 1, tbo, tpe, ktrim = r, ftm = 5 and (2) to remove phiX and Illumina trace contaminants

bbduk.sh was used with parameters $k = 31$, hdist = 1. Illumina adapter reference sequences, the phiX genome and Illumina traces were provided by JGI.

Reads were further trimmed with Sickle (https://github.com/najoshi/sickle) using default settings. Paired end read datasets from each sample were assembled independently from one another using idba_ud under default settings, including the –pre_correction option (*Peng et al., 2012*). For scaffolds greater than 1,000 bp, open reading frames were predicted with Prodigal (*Hyatt et al., 2010*) and functional annotations were determined through similarity searches against the UniProt, UniRef90 (*Suzek et al., 2007*) and KEGG (*Kanehisa et al., 2012*; *Ogata et al., 1999*) databases. tRNAs were predicted for each scaffold using tRNAscan-SE (http://lowelab.ucsc.edu/tRNAscan-SE/). To identify 16S rRNA gene sequences, we searched all assembled scaffolds against the manually curated structural alignment of the 16S rRNA provided with SSU-Align (*Nawrocki, Kolbe & Eddy, 2009*). Coverage values for each scaffold were calculated by read mapping using Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) using default settings. The scaffolds, all associated annotations, and coverage information were then processed and uploaded into ggKbase: http://ggkbase.berkeley.edu/angelo_ncbi_2016/organisms. The sequencing reads have been deposited as "Meadow soil samples from Angelo, CA genome sequencing and assembly": SRA302421—2015-10-05T12:25:59.383. Genomes are currently being processed for submission to NCBI under accession PRJNA297196.

We established a phylogenetic profile for each scaffold by comparing the genes to a database of reference genomes. Assignment of scaffolds >8 kb to genome bins was accomplished using emergent self-organizing maps (ESOM) (Fig. S5). The matrix used in the ESOM was built from a combination of series coverage patterns across samples for each scaffold (ten columns) and tetranucleotide frequency of each scaffold (256 columns). Bins were fine tuned to remove scaffolds classified as wrongly binned based on phylogenetic information or other anomalies using the visualization tools provided by ggKbase. Genome bins were named based on placement in phylogenetically informative gene trees and the overall taxonomic profile of each bin. Bin completeness was evaluated based on the recovery of content of a set of 51 single copy genes for bacteria and 38 single copy genes for archaea using a tool developed in *Probst et al. (2016)*. The phylogenetic signal, in combination with aberrant coverage and/or GC content, was used to identify bin contaminants. Draft quality genomes, defined as genome bins from metagenomes, contain at least 70% of the requisite single copy genes within minimal duplication (a firm cutoff for duplicate genes was not used because some arise due to genes split by scaffolding gaps or contig ends).

Time series analysis

The relative coverage for every scaffold encoding a ribosomal protein S3 gene, thus representing a single strain was determined to indicate the relative abundance of each organism in each sample. Coverage values were normalized to account for differences in sample data size. Values for each species in the same phylum were summed to generate the stacked bar chart presented in Fig. 1 in the 'Results' section.

Proteomics methods

For each soil sample, total proteins were extracted from 10 g of soil using NoviPure® Soil Protein Extraction Kit (MoBio). The crude protein extracts were concentrated to ~1 ml using Amicon® Ultra-4 Centrifugal Filter Units (30 KDa molecular weight cut-off, Millipore). Trichloroacetic acid was then added to precipitate proteins overnight at 4 °C. Proteins were pelleted by centrifugation at 4 °C, washed with ice-cold acetone three times, and re-solubilized in guanidine (6 M) and dithiothreitol (10 mM). Bicinchoninic acid assays were conducted to estimate the protein concentration before adding dithiothreitol. 50μg of proteins from each soil sample was further processed with the filter-aided sample preparation (*Wisniewski et al., 2009*). Proteins were first trypsin digested overnight in an enzyme:substrate ratio of 1:100 (weight:weight) at room temperature with gentle shaking, followed by a secondary digestion for 4 h. All digested peptide samples were stored at −80 °C.

LC-MS/MS proteomic measurements were carried out with 11-step online multidimensional protein identification technology (MudPIT) (*Washburn, Wolters & Yates, 2001*) on an LTQ Orbitrap Elite mass spectrometer (Thermo Scientific), as described previously (*Li et al., 2014*). In each MudPIT run, 25 μg of peptides were loaded offline into a 150-μm-I.D. two-dimensional back column (Polymicro Technologies) packed with 3 cm of C18 reverse phase (RP) resin (Luna, Phenomenex) and 3 cm of strong cation exchange (SCX) resin (Luna, Phenomenex). The back column loaded with peptides was de-salted offline with 100% Solvent A (95% H_2O, 5% CH_3CN and 0.1% formic acid) and washed with a 1-h gradient from 100% Solvent A to 100% Solvent B (30% H_2O, 70% CH_3CN and 0.1% formic acid) to move peptides from RP resin to SCX resin. Then, the back column was connected to a 100-μm-I.D. front column (New Objective) packed in-house with 15cm of C18 RP resin and placed in-line with a U3000 quaternary HPLC pump (Dionex). Each MudPIT run was configured with 11 SCX fractionations using 5%, 7%, 10%, 12%, 15%, 17%, 20%, 25%, 35%, 50% and 100% of Solvent D (500mM ammonium acetate dissolved in Solvent A). Each SCX fraction was separated by a 110-min RP gradient from 100% Solvent A to 50% Solvent B. The LC eluent was directly nanosprayed (Proxeon) into an LTQ Orbitrap Elite mass spectrometer (Thermo Scientific). Both MS scans and HCD MS/MS scans were acquired in Orbitrap with the resolution of 30,000 and 15,000, respectively. The top 10 most abundant precursor ions were selected for MS/MS analysis by HCD after each MS scan. Peptides of each soil sample were measured in technical duplicates.

A protein sequence database was constructed from 3,408,250 full-length predicted proteins (combined file size of 863.56 Gb) by metagenomics from four samples; (1) pre-rain plot 1 10–20 cm, 2 days after second rain (2) plot 1 10–20 cm, (3) plot 1 30–40 cm, and (4) plot 2 10–20 cm, and their reverse sequences as decoys for estimation of false discovery rate (FDR) (*Elias & Gygi, 2007*). Database searching was performed with Sipros 3.0 (*Hyatt & Pan, 2012*; *Wang et al., 2013*) on the Titan supercomputer at Oak Ridge Leadership Computing Facility. The following parameters were used: dynamic oxidation of methionine, static alkylation of cysteine by iodoacetamide, 0.03 Da mass tolerance for precursor ions and 0.01 Da for fragment ions, up to three missed cleavages, and full

enzyme specificity required. The FDR was strictly controlled at the peptide level (1%). One unique peptide was required for each identified protein/protein group. Indistinguishable proteins were combined into protein groups based on the parsimony rule (*Nesvizhskii & Aebersold, 2005*). The numbers of identified protein/protein group per sample ranges is provided in Table S3B. Proteins were linked to draft genomes so the functions could be assigned to individual organisms. Spectral counts of proteins were normalized across the samples for label-free quantification as described previously (*Pan & Banfield, 2014*; *Wang et al., 2013*). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (*Vizcaíno et al., 2016*) partner repository with the dataset identifier PXD004965.

Metabolomics methods

Metabolites were extracted from sieved (2 mm) soil samples using an aqueous extraction protocol (modified from *Swenson et al., 2015*). Briefly, sieved soils (4 g) were incubated (200 rpm, 1 h, 4 °C) in triplicate with MilliQ water (16 ml) amended with 1.6 μg/ml of ABMBA (2-Amino-3-Bromo-5-methylbenzoic acid) and 2 μg/ml of UL-13C-Glucose included as internal extraction standards. Samples were centrifuged (3,220 × g, 15 min, 4 °C) and the supernatant was carefully decanted. Soils were then back extracted twice with MilliQ water (4 ml, 2 ml) as above but with abbreviated incubation times (15 min and 30 s). Following centrifugation, supernatants were pooled, frozen (−80 °C) and lyophilized. Metabolites were then re-suspended in 100% methanol (250 ml, to limit salt solubility) with internal standards (5 μg/ml—Table S5) and filtered (Millipore, 0.22 mm PVDF microcentrifuge filters). Samples were analyzed in random order on an Agilent 6550 iFunnel Q-TOF LC/MS system with a SeQuant ZIC-pHILIC zwitterionic exchange column (150 × 2.1 mm, 5 mm, Merck Millipore) using a neutral pH (5 mM NH_4OAc)/acetonitrile gradient (10% to 55% aqueous phase over 25 min at 0.25 ml/min) in both positive and negative ionization mode. Metabolites and putative metabolites were identified manually through several methods including: comparison to the Northen Lab standards library (>290 compounds), MS/MS analyses, and formula generation (Agilent MassHunter) (Table S5). Peak detection and areas were determined using the Northen Lab's Metabolite Atlas software (http://metatlas.nersc.gov) (Table S6). Internal standards (including extraction standards) were used for quality control analyses and to detect and control for retention time shifts and other analytical variability in any of the sample analyses. A one-way ANOVA was conducted on metabolites across the samples to determine if they changed significantly between samples; *p*-values were then corrected for multiple comparisons (*Benjamini & Hochberg, 1995*) (Table S7). Significant changes were determined for metabolites between sampling times at each depth and between depths in each sample using ANOVA with post-hoc Tukey HSD pairwise analyses (Table S8).

16S rRNA and rpS3 phylogenetic analyses

For the 16S ribosomal RNA (rRNA) tree and ribosomal protein S3 (rpS3), alignments were generated from all 16S rRNA and rpS3 genes available the metagenomes. Sequences are provided in the Supplemental Information. All 16S rRNA genes longer than 660

bp were aligned using the SINA alignment algorithm through the SILVA web interface (*Pruesse, Peplies & Glöckner, 2012*; *Pruesse et al., 2007*). All rpS3 amino acid sequences longer than 180 aa long were aligned using MUSCLE (*Edgar, 2004a*; *Edgar, 2004b*). The full alignments were stripped of columns containing 95% or more gaps. A maximum-likelihood phylogeny was inferred using RAxML (*Stamatakis, 2014*) run using the GTRCAT model of evolution for the 16S rRNA and PROTGAMMLG for the rpS3. The RAxML inference included calculation of 300 bootstrap iterations for the 16S rRNA tree and 100 for the rpS3 tree (MRE-based Bootstopping criterion), with 100 randomly sampled to determine support values.

Ordination analyses of microbial community structure

Ribosomal protein S3 genes were retrieved using HMMs build from the dataset published in *Hug et al. (2015)* and used to search against predicted protein sequences in all samples. Only sequences that spanned at least 60% of the alignment were included and clustered at 99% similarity (equivalent to species level *Sharon et al., 2015*) using Usearch (clusterfast, *Edgar (2010)*). Abundances of each cluster in each sample were determined from scaffold coverage (see above) and normalized to percent abundances in each sample. Principle coordinate analysis (PCoA) based on Bray-Curtis distance measure was computed using the R programming environment (*R Core Team, 2015*) in conjunction with the vegan package (*Dixon, 2003*).

Methanol dehydrogenase tree

The amino acid sequences of the PQQ-dependent methanol dehydrogenase proteins detected in the proteomics data were aligned to reference sequences with MUSCLE and this alignment was used to build a tree with RAxML with the PROTGAMMAWAG model and 100 bootstrap iterations.

Physical and chemical characterization of the soil

Samples of soil and the weathered bedrock (mudstone and sandstone) were collected for mineralogical and other analyses using electron microprobe, X-ray diffraction and scanning electron microscopic methods. Minerals identified included vermiculite (the predominant mineral in the soil), plagioclase and alkali feldspars, minor apatite and a mixture of Fe, Mn-oxyhydroxides.

RESULTS

Between 14 and 25 Gb of DNA sequence data were obtained per soil sample for the ten soil samples, two of which were time and depth replicates. The 250 bp reads were assembled as detailed in 'Methods,' resulting in 2,982,775 contigs > 1,000 bp in length (42,770 contigs > 10,000 bp, the longest was 538,000 bp). In total, these contigs encode 8,773,880 genes (Table S1). For each sample, the reads were then mapped back to the assembled contigs of the sample to generate coverage statistics.

We compared the microbial community structure across the sample series using coverage of scaffolds assembled from the sequence datasets. Out of the 1,420 microorganisms

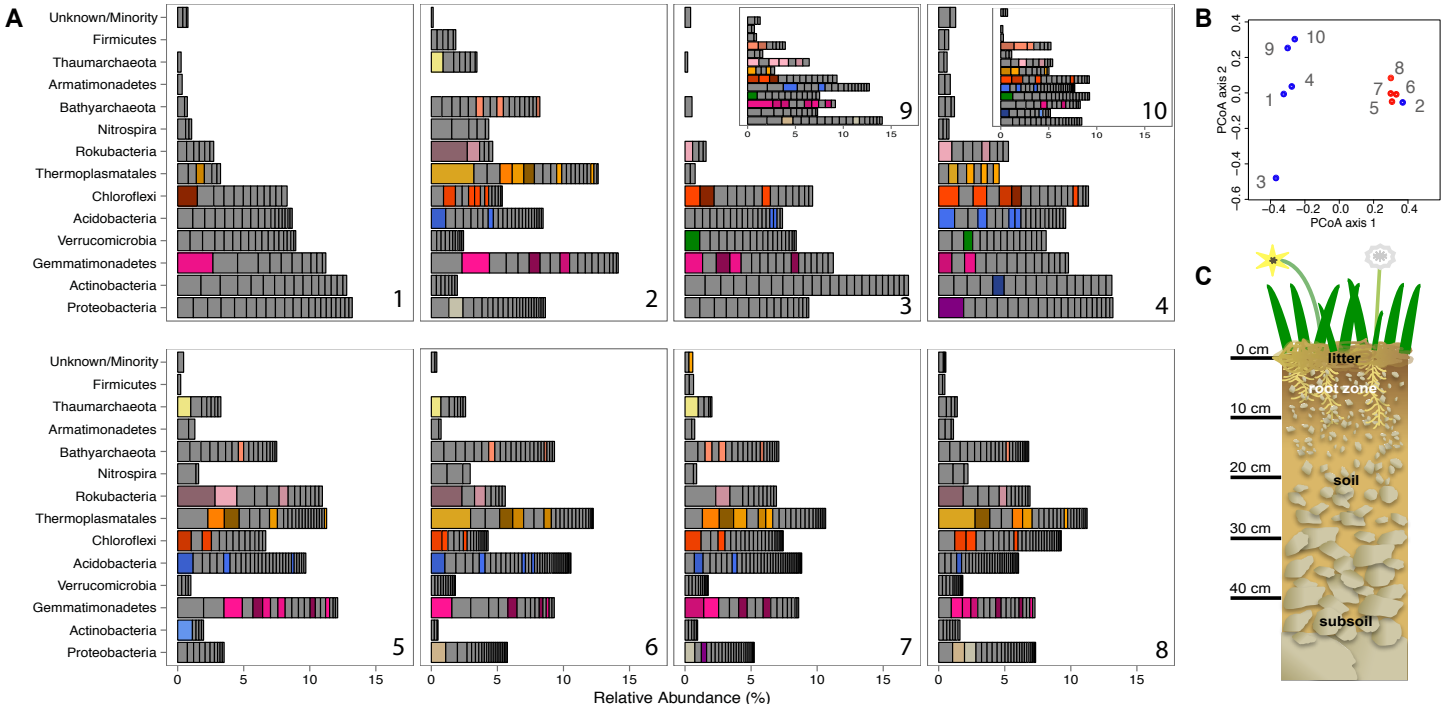


Figure 1 **Prokaryotic diversity and abundance over sampling time and depth.** Ribosomal protein S3-encoding scaffold relative abundances for each sample are plotted in (A), and organized as follows: pre-rain (1, 5), four days after the first rainfall (2, 6), six days after the first rainfall (3, 7, 9) and two days after the second rainfall (4, 8, 10) at 10–20 cm (1–4, 9, 10) and 30–40 cm deep (5–8). Persistent and abundant species (>1% abundant in multiple samples) are colored by phylum and shaded by species. The beta diversity PCoA plot (B) compares the ten sample communities (blue denotes 10–20 cm and red denotes 30–40 cm samples). A diagram depicting the sampled ecosystem (C) shows the shallow roots of the annual forbs and grasses mostly remain in the top 10 cm and the rocks increase in size in the deeper soil.

detected based on marker gene (rpS3) sequences, 652 occurred in 2–18 times (average: 4.2 ± 2.86) over the ten samples and had relative abundances of between 0.06 and 3.2% of the community (Fig. 1). The same information was used in an ordination analysis and showed that samples from the same depth were more similar to each other than to those from the other depth with the exception of the first post-rain 10–20 cm sample (Fig. 1). The results indicate substantial overlap in the organisms present, especially among samples from the same soil depth. The same organisms were also observed in soil samples collected at the same depth and time from sites separated by ∼10 m. In fact, many of these organisms are highly abundant, ranking in the top ten most abundant organisms and amounting for > 1% relative abundance in each sample (Fig. 1). The availability of sequence information for multiple independently assembled samples with substantial overlap in community membership enabled the addition of series abundance parameterization to nucleotide frequency-based genome binning. We generated 198 bins, 46 of which were classified as metagenome-assembled draft genomes. Most sequence fragments that could not be assigned to specific genomes were grouped at the phylum-level and assigned to "megabins" (Table S2A). However, most of the relatively high coverage scaffolds in each sample were binned into draft genomes.

A notable feature of the soil microbial community compositions, evident based on phylogenetic analysis of single copy genes, is the high representation of organisms in the sub-root zone soils from phyla that are relatively poorly represented in the NCBI database. For example, we obtained draft genomes for two Gemmatimonadetes, two Verrucomicrobia, eight Acidobacteria, one Armatimonadetes, three Chloroflexi, and three Nitrospirae. Importantly, we also reconstructed seven draft genomes from the Rokubacteria, a bacterial Candidate Phylum first reported from aquifer sediment in 2015 (*Hug et al., 2016*). In addition, we reconstructed draft genomes for one Betaproteobacterium, two Deltaproteobacteria an Actinomycetales and one novel Actinobacteria. The reconstructed genomes, which represent all major phyla detected in the soil samples, substantially expand the coverage of phylogenetic diversity in the NCBI database (Table S2B). In addition, we generated four draft genomes from the "Miscellaneous Crenarchaeota Group" (MCG), two from the "Soil Crenarchaeota Group" (SCG) and three from the "South African Gold Mine Miscellaneous Group" (SAGMCG). The SCG and SAGMCG are likely within the Thaumarchaeotes whereas the MCG are novel Bathyarchaeota (Figs. S2A and S2B). We also sampled Euryarchaeotes from the Thermoplasmatales lineage, but the genome bins were not well resolved.

Phylogenetic trees constructed using marker genes for both bacteria and the archaea from samples collected from different depths and times display structures similar to the seed puff of a dandelion, with many closely related strains at the termini of most branches (Fig. 2 and Fig. S2). In prior studies, strains were grouped at higher taxonomic levels, up to the phylum level (*Delmont et al., 2015*; *Hultman et al., 2015*). The Gemmatimonadetes phylum branch of the ribosomal protein S3 phylogenetic tree (Fig. 2) exhibits fine scale diversity that is comparable to the observed level of diversity detected in every phylum. Many organisms of the same type (separated by zero branch length in Fig. 2) occurred in samples collected at different times and from both depths.

Before the rain, the most abundant organisms in the shallow depth interval (10–20 cm) are Gemmatimonadetes and Actinobacteria species. Gemmatimonadetes species are also abundant after the rain event. Thermoplasmatales are highly abundant in samples collected after the rain event (Fig. 1).

In general, the microbial community composition of 30–40 cm depth samples differed from that of the 10–20 cm depth samples and samples collected before and after the rain event were less different than those collected from the shallower soil. Notably, the community sampled from deeper soil prior to the rain event was dominated by a Rokubacterium and several Thermoplasmatales (Euryarchaeota) were abundant. In fact, around 20% of the microorganisms sampled in the deeper soil interval were archaea, an interesting finding because archaea are generally believed to be relatively rare compared to bacteria in soil (*Fierer et al., 2012*). In samples collected after the rain, three Thermoplasmatales archaea are highly abundant and rank in the top 10 most abundant organisms in the deeper soil (Fig. 1).

We obtained proteomic information to provide insight into the active pathways that mediated carbon and nitrogen compound transformations. Between 2,881 and 4,716 proteins/protein groups were identified per soil sample (Table S3A). Overall, we identified

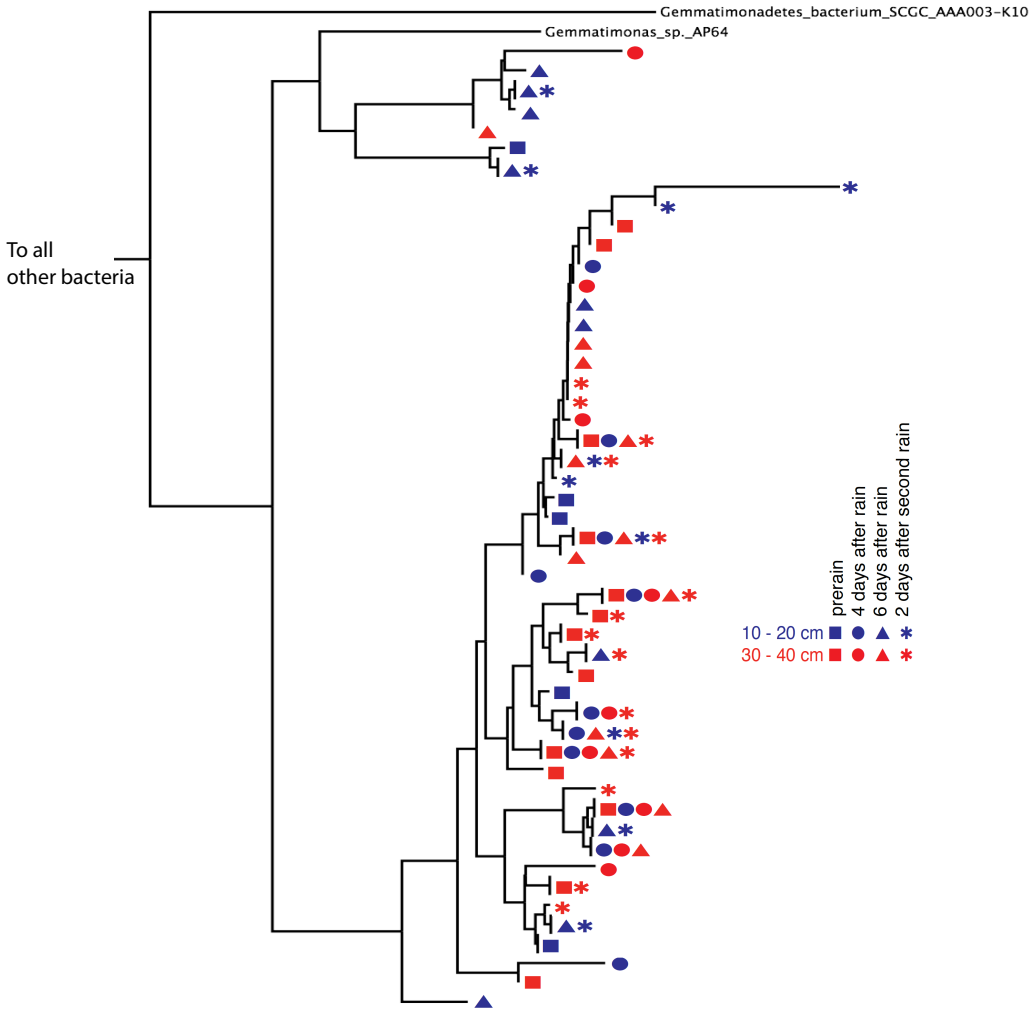


Figure 2 Gemmatimonadetes phylogenetic ribosomal protein S3 tree. Subsection of the experiment's ribosomal protein S3 phylogenetic tree shows typical diversity within the phyla: dozens of novel, closely related organisms of the Gemmatimonadetes bacterial phylum inhabit the two soil zones (10–20 cm, blue and 30–40 cm, red) before (squares) and after the rain events (four days after: circles, six days after: triangles, and two days after the second rain: asterisks). Identical ribosomal protein S3 sequence branches representing members of the same species were collapsed and the time and depth symbols of these members are presented horizontally.

6,835 proteins and 6,378 protein groups in the soil microbiome, based on 28,782 distinct peptide identifications. It is important to note that we could link most identified proteins to the specific microorganisms from which they derived because a significant fraction of our sequence data was genome-resolved.

Significantly, the most abundant protein in the proteome was not involved in plant sugar breakdown. Rather, it was pyrrolo-quinoline quinone (PQQ)-dependent methanol dehydrogenase (MDH). This protein is encoded in the genome of a Gemmatimonadetes bacterium. PQQ dependent MDH (PQQ MDH) is the second step of methanotrophy, and follows the oxidation of methane to methanol by methane monooxygenase. However, methane monooxygenase was not present in any of the genomes, including those harboring

the methanol dehydrogenase. Thus, we infer that this enzyme is involved in methylotrophy, the consumption of methanol. Lesser abundant PQQ MDH proteins from Rokubacteria were also detected in the proteomics analysis

MDH proteins have typically well-conserved sequences. The MDH proteins represented in the proteome were aligned with sequences from the literature (*Taubert et al., 2015*) and a tree was built, resulting in the clustering of the proteins by phylum. The Rokubacteria XoxF sequences formed a distinct branch in the MDH protein tree (XoxF and MxaF families) and the Gemmatimonadetes sequences formed a new clade with the XoxF sequences from various Proteobacteria (*Taubert et al., 2015*) (Fig. 3) The Gemmatimonadetes PQQ-MDH proteins were generally more abundant than those from Rokubacteria (as shown in the heat map on the right side of Fig. 3). The MDH proteins all contain the catalytic and cofactor binding residues required for activity (*Anthony & Williams, 2003*), including those for PQQ, as well as the aspartate residues thought to select for the lanthanides such as Ce^{3+} and La^{3+} (*Keltjens et al., 2014*; *Pol et al., 2014*). The selection for lanthanides over Ca^{2+} is an interesting bioinorganic trait because while lanthanides are abundant in the Earth's crust, they are highly insoluble and thus considered biologically unavailable and in turn not well studied (*Skovran & Gomez, 2015*). Despite this, La^{3+} is required for the activity of the XoxF type MDH perhaps because it a more efficient Lewis acid in the polarization of PQQ than Ca^{2+} (*Bogart, Lewis & Schelter, 2015*; *Pol et al., 2014*).

Following oxidation of methanol by PQQ MDH the toxic formaldehyde product must be moved from the periplasm to the cytosol for transformation and incorporation into 3C compounds. The reaction can occur via one of three pathways: the glutathione, tetrahydromethanopterin (THMPT), or tetrahydrofolate (THF)-linked formaldehyde oxidation pathways (*Chistoserdova, 2011*). The Gemmatimonadetes genomes have the entire THMPT formaldehyde oxidation pathway and the tetrahydrofolate to serine pathway for these reactions. Furthermore, they encode the PQQ biosynthesis machinery. The THMPT biosynthesis machinery is encoded immediately upstream of the PQQ MDH gene (Fig. S3) (*Scott & Rasche, 2002*). The results strongly support the capacity for methylotrophy in these soil-associated Gemmatimonadetes and functioning of this pathway (Fig. 4 and Fig. S4).

Although the Rokubacteria have PQQ MDH, the genomes do not encode any of the three pathways for formaldehyde transformation. It has been suggested that the XoxF type of MDH is able to convert methanol directly to formate, bypassing the formaldehyde oxidation mechanism (*Pol et al., 2014*). Formate then may be broken down using formate dehydrogenase to yield energy or be assimilated by one of several pathways. The Rokubacteria also may be carrying out beta-oxidation of fatty acids, as these proteins for this pathway are abundant in the proteome.

We identified many proteins likely responsible for decomposing matter from the meadow's early summer senescing annual grasses (*Bromus hordeaceus, Bromus diandrus, and Bromus tectorum*) and annual lupine (*Lupinus bicolor*) (for a full species list, see Table S4). Highly represented were carbohydrate-active enzymes such as glycoside hydrolases, polysaccharide lyases, and many sugar and amino acid transport proteins (Table S3A). Notably, enzymes of the Thermoplasmatales and Bathyarchaeota archaea

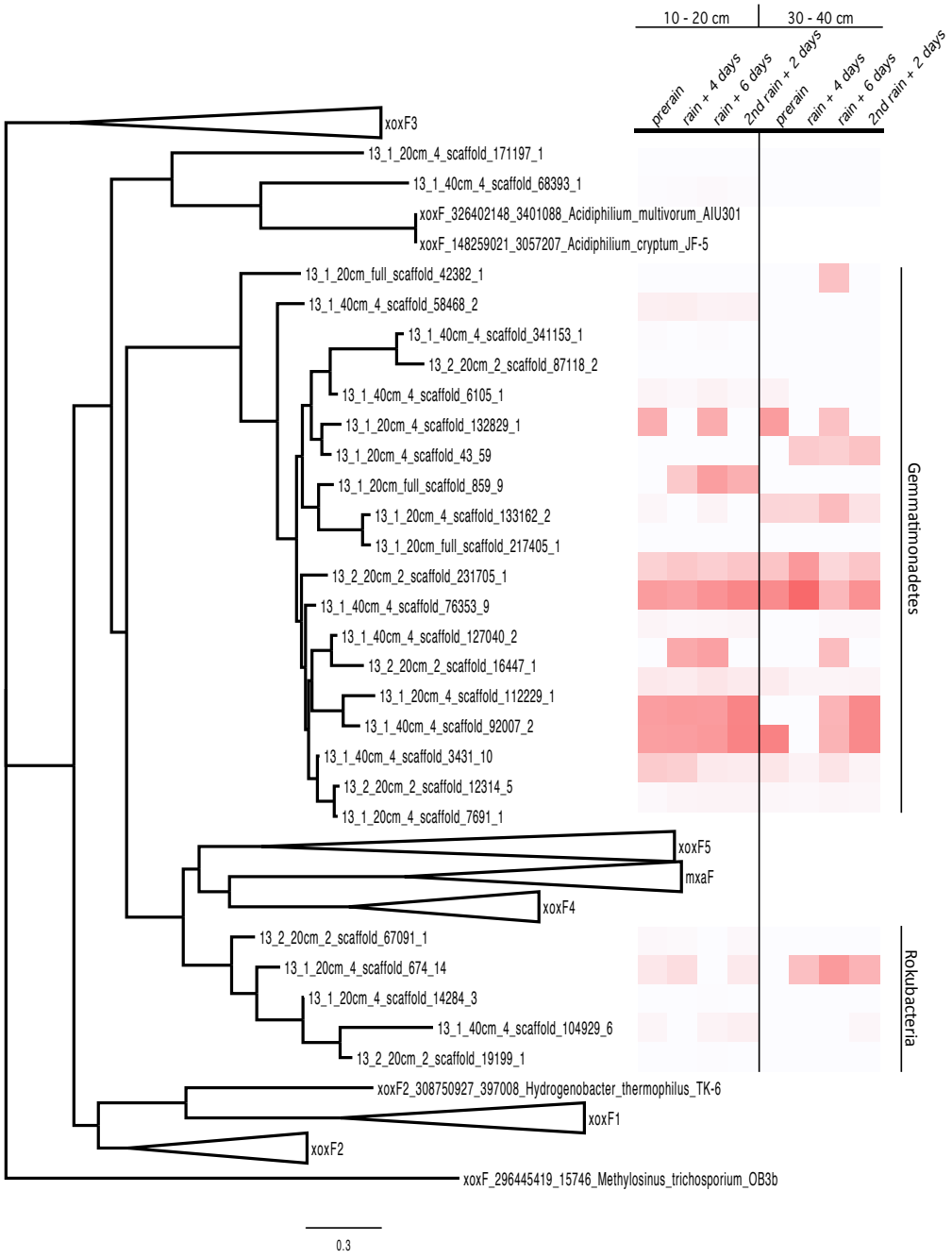


Figure 3 PQQ-dependent methanol dehydrogenase (XoxF and MxaF) protein clades and abundances in the soil zones. PQQ-dependent methanol dehydrogenase (XoxF and MxaF) protein tree containing sequences from the literature and experimental sequences in the soil zones with their corresponding relative abundance of normalized spectral counts from the proteomics results in a heat map.

for protein uptake and degradation, such as extracellular serine-type endopeptidases (annotated as pyrolisin-like serine protease and encoded with a N-terminal signal peptide), amino acid transporters, and cytoplasmic amidases and formamidases highly represented in the proteome. The findings parallel results for marine Thermoplasmatales

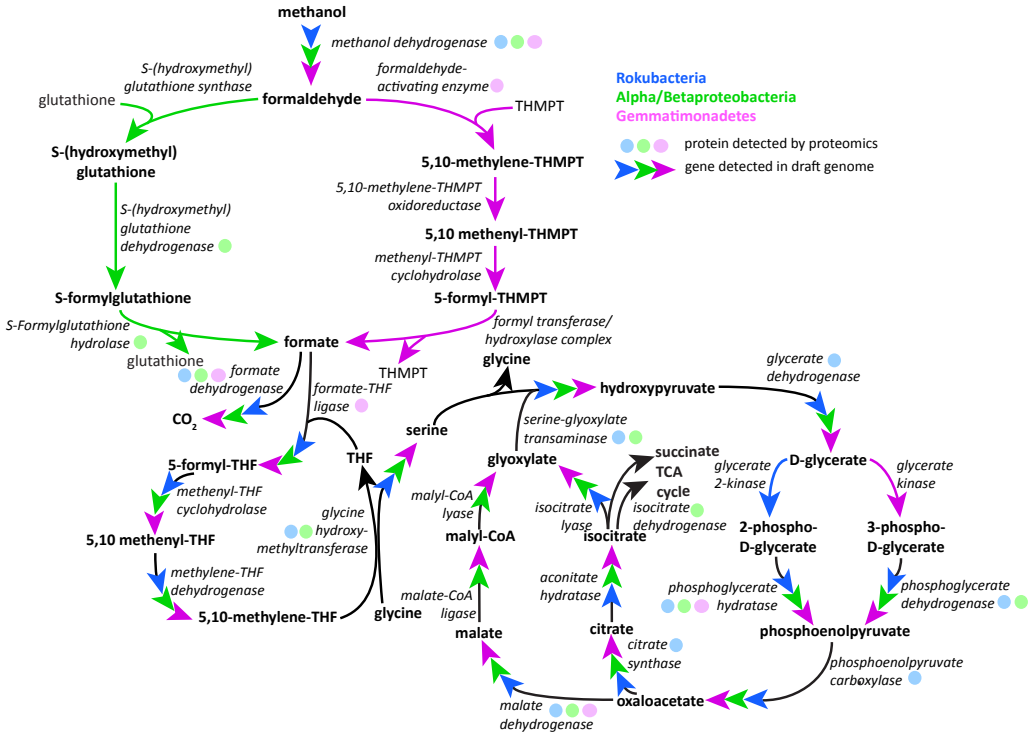


Figure 4 Putative active methylotrophy pathways in Rokubacteria, Proteobacteria, and Gemmatimonadetes. The colored arrows indicate genetic evidence of proteins that catalyze the reactions that transform methanol into CO_2 or cell material and colored dots indicate representation in the proteome. Gemmatimonadetes (purple) and Proteobacteria can oxidize formaldehyde via the tetrahydromethanopterin and glutathione-linked pathways, respectively, while Rokubacteria (blue) cannot. Formate is incorporated by the tetrahydrofolate-linked pathway then converted to serine by glycine hydroxymethyltransferase. Serine and glyoxylate are used by serine-glyoxylate transaminase to produce glycine and hydroxypyruvate, which is then converted to D-glycerate. D-glycerate can be phosphorylated by two different kinases, glycerate 2-kinase in Rokubacteria and glycerate kinase in Gemmatimonadetes. Both phosphoglycerates are converted by different enzymes to phosphoenolpyruvate (phosphoglycerate hydratase or phosphoglycerate dehydrogenase) and then to oxaloacetate by phosphoenolpyruvate carboxylase. Gemmatimonadetes and Proteobacteria (green) contain but Rokubacteria lack the canonical malate-intermediated serine formaldehyde assimilation pathway but do encode citric acid/glyoxylate cycle genes that could assimilate carbon (citrate synthase) and regenerate glyoxylate (isocitrate lyase).

(*Lloyd et al., 2013*) and members of the TACK superfamily from groundwater (*Castelle et al., 2015*), and suggest a role for novel soil Archaea in protein degradation.

We investigated the distribution of metabolites in triplicate extracted soil water samples using a robust Liquid Chromatography-based Mass Spectroscopy (LC-MS) metabolomics workflow based on a previous study on samples from the meadow soils (*Swenson et al., 2015*). A total of 125 unique compounds were detected and quantified. These include sugars (one to six sugar residues per chain), sugar alcohols, amino acids, nucleotides/nucleosides, quaternary amines, osmolytes and several suspected sugar metabolites and derivatives (Fig. 5). Notably, most concentrations fall to zero at the base of the soil zone, an observation that suggests efficient scavenging of these compounds by microbial soil communities and not the sorption to mineral surfaces (*Fischer, Ingwersen & Kuzyakov, 2010*).

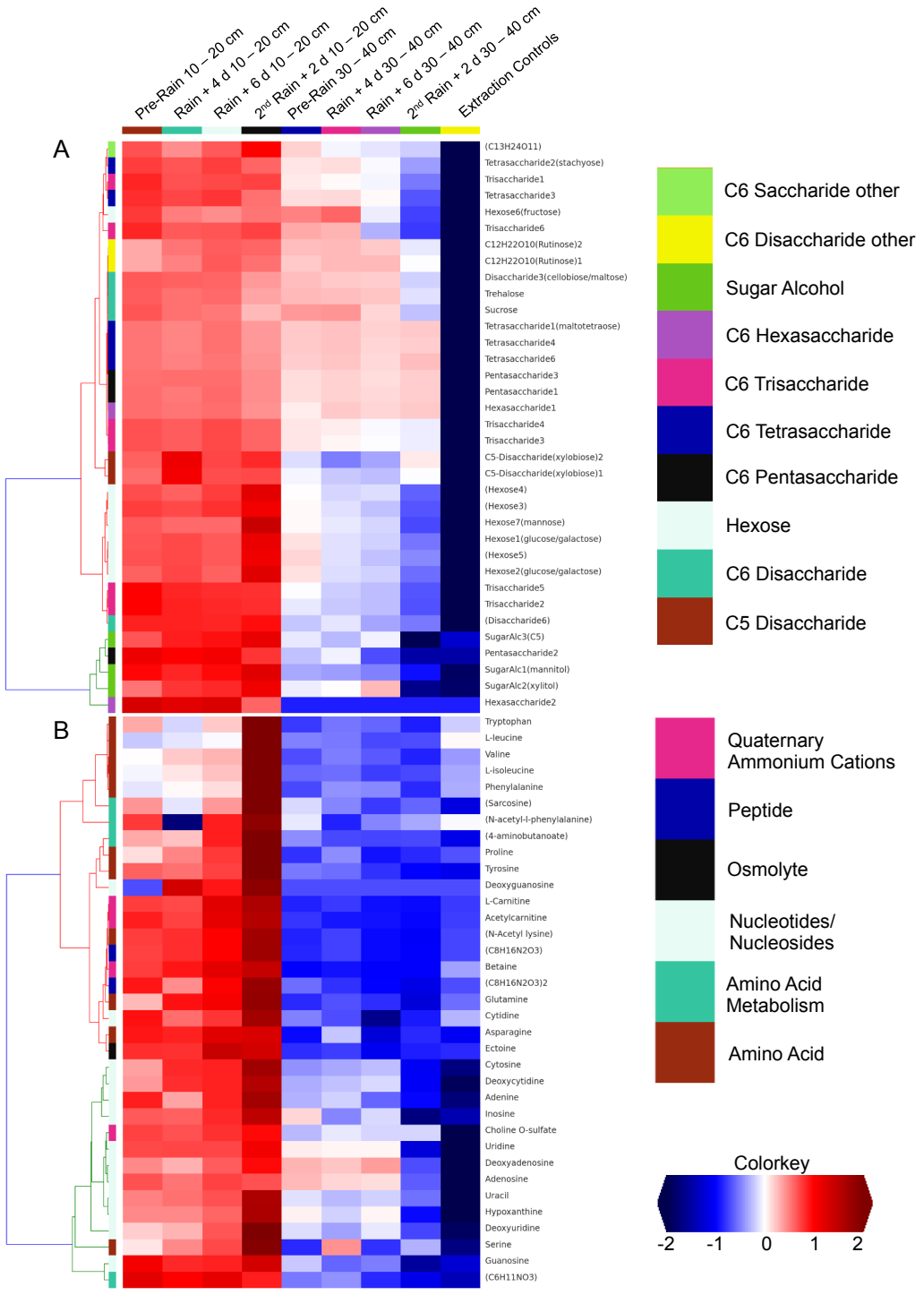


Figure 5 Comparison of detected sugars and common nitrogenous compounds in Angelo soils at 10–20 cm and 30–40 cm. (A) lists sugars and (B) lists nitrogen-containing compounds and the heat map indicates relative mean sample concentrations. Columns are ordered left to right by date. Row colors (left sides of plot) are based on chosen metabolite groupings. A clear decrease in metabolite abundances is observed with increasing soil depth.

Dissolved organic nitrogen, which is comprised of nitrogen-bearing molecules such as amino acids, represents an important nutrient source for soil microbial communities (*Jones et al., 2004*). The metabolomic data identifies dramatic increases in amino acid and nucleotide concentrations at 10–20 cm right after the second rain event (Fig. 5). The timing of this increase, immediately following a substantial rainfall event, points to the vertical transport of these compounds from the top 10 cm. However, the source and localization of their production were not determined in the current experiment. Notably, there was no observed accumulation of these compounds in the deeper soil (30–40 cm). This may be attributable to microbial activity at these depths. For example, the Thermoplasmatales and Bathyarchaeota archaea, which were highly represented in the 30–40 cm depth soil especially after rain events, have proteins for uptake and aerobic breakdown of peptides. Abundances of these proteins, which include dipeptide, oligopeptide, hydrophobic peptide, polar peptide transporters are high in samples collected after the first rainfall (Table S3A).

Interestingly, ammonia-producing formamidase is the most abundant protein identified for several archaea. Ammonia liberated by archaeal peptide degradation likely supports growth of SAGMCG and SCG Thaumarchaeota, which both have ammonia monooxygenases genes. A copper-containing nitrite reductase (NirK), along with cytochromes, sulfurtransferases and Fe-S proteins, were abundant in the proteome but NO reductases were not identified by proteomics (although they are encoded in the genomes). Overall the results suggest roles for Thaumarchaeota in both nitrification and denitrification

Breakdown of plant-derived organics can release sulfur compounds. For example, glucosinolates are sulfur-bearing organics that are produced by Brassicales, a widely distributed group of plants in the mustard family that were identified in this study's meadow. Degradation products of glucosinolates include sulfur-containing thiocyanate and isothionates. The Rokubacteria have genes of the Sox sulfur oxidation pathway, as do the first-described Rokubacteria described from groundwater (*Hug et al., 2016*), and some Sox proteins were identified in the proteome. Thus, these novel bacteria are inferred to play an important role in sulfur biogeochemistry in the sub-root zone soil during the rainfall-induced period of organic matter turnover.

Choline sulfate is an interesting sulfur-containing osmolyte identified in the metabolome. This compound is produced by many organisms, including plants and fungi, and is degraded to produce betaine via sulfatase enzymes. Betaine was also identified by metabolomics. Notably, sulfatases were observed in the proteomes of Actinobacteria, Chloroflexi, Alphaproteobacteria and Betaproteobacteria. Sulfate released from choline sulfate degradation may play an important role in soil bacterial sulfur metabolism (*Markham et al., 1993*). Choline sulfate and betaine both contain three methyl groups per molecule that are released upon the degradation to glycine and could contribute to the growth of the methylotrophic bacteria.

DISCUSSION

Our study design aimed to deeply analyze soil microbial community composition and to detect genes and proteins from many organisms, including those at relatively low

abundance. This generated a soil metagenomic dataset of unprecedented size (>200 Gb) and complexity (>1,400 species). Given the massive data sizes involved in this research, we limited analysis to ten samples so we do not attempt to describe overall shifts in the patterns of microbial distribution during this time period. However, the spectrum of conditions provided access to a wider variety of genomes than would be provided if analyses target a single sampling location or time point. Also, we extracted DNA from large (~200 g) homogenized samples; this likely reduced the impact of spatial microheterogeneity and probably explains why overlap in community composition could be detected.

Only recently has genome reconstruction from soil (seven from permafrost by *Hultman et al. (2015)*, seventeen from enrichments by *Delmont et al. (2015)*, and 129 from prairie soil by *White et al. (2016)*) been achieved. In our study, no single organism represented >3% of the community. The extensive strain and within population variation likely explains the high level of fragmentation of genomes for some organisms (Tables S1 and S2). Even with these challenges, we reconstructed genomes from all the major lineages represented in the microbial communities. We attribute this result to the assembly of reads into large scaffolds, many of which could be binned because overlap in community composition over the sample series enabled binning using abundance pattern information. Scaffold assembly provided high quality ribosomal protein S3 sequences that were used to distinguish organisms at the species level, a phylogenetic resolution exceeding that which could be obtained by rRNA sequencing methods (*Sharon et al., 2015*). The predicted protein dataset provided the foundation for multi-omic analyses that yielded functional insights.

Most prior research on carbon cycling in soil has focused on microbial degradation of complex soil organic macromolecules, likely derived in part from plant biomass. Our metabolomics analysis suggest that the organic and nitrogenous substrates needed to sustain microbial life disappear from the soils relatively rapidly as few metabolites accumulate to measurable amounts in the deeper soil profile. Nitrogenous compounds, mostly free amino acids, were identified in soil after the second rain in the 10–20 cm zone yet were practically undetectable in the 30–40 cm zone. It is clear that the substrate availability between one 10 cm zone to the next is very different to that in the upper soil horizon, and will support different microbes. For example, the Verrucomicrobia and Actinobacteria are only abundant in the 10–20 cm depth interval and much rarer in the 30–40 cm depth interval. We also found abundant, diverse proteins involved amino acid and carbohydrate degradation and import in every partial to near-complete draft genome. Because we employed an untargeted proteomics, we could identify many thousands of proteins using a peptide database composed of full-length genes predicted from the samples' metagenomes.

Yet, interestingly the most abundant protein in the proteomics data was the PQQ-dependent methanol dehydrogenase from Gemmatimonadetes and Rokubacteria. In addition to complex carbohydrates, plant biomass and root exudates also provide an abundant source of methanol (*Sutton & Sposito, 2005*). Previously, only members of Proteobacteria, NC10, and Verrucomicrobia have been shown to be methylotrophs (*Chistoserdova, 2011*; *Op den Camp et al., 2009*). Methylotrophy was tentatively linked to Gemmatimonadetes only once before, when ^{13}C methanol containing compounds were

fed to a lake sediment sample and labeled Gemmatimonadetes 16S rRNA was identified (*Nercessian et al., 2005*). Methylotrophy has been described in aerobic lake sediments (*Costello & Lidstrom, 1999*), the phyllosphere (*Corpe & Rheem, 1989*; *Delmotte et al., 2009*) marine (*Radajewski et al., 2002*; *Stacheter et al., 2013*), and soil (*Eyice et al., 2015*; *Kolb, 2009*; *Radajewski et al., 2002*; *Stacheter et al., 2013*) environments. These studies found that methanol-oxidizing enzymes of Proteobacteria have micro- and nanomolar affinity for methanol, the highest activity occurring in the root-associated soil, and that methylotrophic communities thrive under the full range of plant diversity and soil pH (*Radajewski et al., 2002*; *Stacheter et al., 2013*). Further, methylotrophic methanogenesis can occur under aerobic conditions (*Hofmann et al., 2016*; *Karl et al., 2008*; *Metcalf et al., 2012*). However, in our study, no methyl-coenzyme M reductase complex (*mcrA*) gene was predicted in any dataset. Thus, methylotrophy is neither occurring in nor linked to co-occurring methanogens.

Notably, we observe a significant fraction of the microbial community (87 distinct organisms via rpS3 genes) belong to the as yet uncultured yet widespread phylum Bathyarchaeota (formerly known as the Miscellaneous Crenarchaeotal Group) (*Gagen et al., 2013*; *Kubo et al., 2012*). Bathyarchaeota have been identified by 16S rRNA studies of sulfate-methane transition zones and hypothesized as being involved in dissimilatory anaerobic methane oxidation coupled to organic carbon assimilation (*Biddle et al., 2006*). In a recent report, Bathyarchaeota were identified in metagenomic analyses of coal-bed methane well water. The genomes encoded a complete methanogenic pathway including an ancient *mcrA* (*Evans et al., 2015*). The four draft (71–91%) Bathyarchaeota genomes do not contain the genes required for methanotrophy or methanogenesis but encode oligopeptide import and amino acid degradation pathways, which were also abundant in the proteomics analysis. The large transporter diversity suggests substantial substrate flexibility in Bathyarchaeota (and also in Thermoplasmatales archaea). Thus, along with other community members including Thermoplasmatales, Bathyarchaeota likely contribute to degradation of nitrogen-containing compounds in the deeper soil. The findings underline the importance of genomic resolution, because metabolic roles of the soil Bathyarchaeota predicted based on phylogenetic information and previously published genomes would have been incorrect.

CONCLUSION

This genome-resolved multi-omic study revealed many populations of little known bacteria and archaea in sub-root zone soil microbial communities. Our proteogenomic analysis yielded strong evidence for methanol oxidation in novel members of the Gemmatimonadetes and Rokubacteria phyla. These capacities have not been previously linked to organisms of these phyla, although Gemmatimonadetes are common members of soil microbial communities. Rokubacteria, on the other hand, have not previously been reported from soil, so the findings of this study contribute new information regarding microbial community composition as well as function. Removal of methanol and other small organic molecules from solutions draining from upper soil horizons by these

bacteria limits their availability for metabolism by organisms at greater distance from leaf-litter associated carbon sources. Although methanogeneis is not prominent in the studied grassland, such activities in other soils could restrict the supply of methanol to methylotrophic methanogens in deeper subsurface regions. We found that different microbes and metabolites are abundant in samples collected just 10 cm apart. Likely, the organisms are stratified by substrate availability, a pattern that results in part from the activities of organisms in the overlying soil regions.

ACKNOWLEDGEMENTS

We would like to thank the rest of the members (and former members) of the Banfield lab for their help with and the development of various tools and reference libraries, and Dr. David Burstein for help with collecting soil samples. The sequencing was conducted by the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, and Lawrence Berkeley National Laboratory.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work is supported by the Office of Science, Office of Biological and Environmental Research, of the US Department of Energy Grant DOE-SC10010566. The sequencing was conducted by the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, and Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Office of Science, Office of Biological and Environmental Research, of the US Department of Energy: DOE-SC10010566.
US Department of Energy Joint Genome Institute.
DOE Office of Science User Facility. Lawrence Berkeley National Laboratory: DE-AC02-05CH11231.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Cristina N. Butterfield, Zhou Li and Peter F. Andeer conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Susan Spaulding performed the experiments, contributed reagents/materials/analysis tools, prepared figures and/or tables.

- Brian C. Thomas performed the experiments, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Andrea Singh performed the experiments, contributed reagents/materials/analysis tools.
- Robert L. Hettich, Trent Northen and Chongle Pan conceived and designed the experiments, contributed reagents/materials/analysis tools, wrote the paper, reviewed drafts of the paper.
- Kenwyn B. Suttle performed the experiments, contributed reagents/materials/analysis tools, wrote the paper, reviewed drafts of the paper.
- Alexander J. Probst performed the experiments, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables.
- Susannah G. Tringe contributed reagents/materials/analysis tools, wrote the paper, reviewed drafts of the paper.
- Jillian F. Banfield conceived and designed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

Field Study Permissions

The following information was supplied relating to field study approvals (i.e., approving body and any reference numbers):

Angelo Coast Range Reserve Permission APP#27790.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

Sequencing reads: ''Meadow soil samples from Angelo, CA genome sequencing and assembly'': SRA302421; Soil metagenome, BioProject PRJNA297196; and all data is also available in ggKbase: http://ggkbase.berkeley.edu/angelo_ncbi_2016/organisms.

Data Availability

The following information was supplied regarding data availability:

The raw data has been supplied as a Supplemental File.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.2687#supplemental-information.

REFERENCES

Aanderud ZT, Jones SE, Schoolmaster DR, Fierer N, Lennon JT. 2013. Sensitivity of soil respiration and microbial communities to altered snowfall. *Soil Biology and Biochemistry* **57**:217–227 DOI 10.1016/j.soilbio.2012.07.022.

Adair KL, Wratten S, Lear G. 2013. Soil phosphorus depletion and shifts in plant communities change bacterial community structure in a long-term grassland management trial. *Environmental Microbiology Reports* **5**:404–413 DOI 10.1111/1758-2229.12049.

Aerts R, Bakker C, De Caluwe H. 1992. Root turnover as determinant of the cycling of C, N, and P in a dry heathland ecosystem. *Biogeochemistry* **15**:175–190 DOI 10.1007/BF00002935.

Anthony C, Williams P. 2003. The structure and mechanism of methanol dehydrogenase. *Biochimica et Biophysica Acta—Proteins and Proteomics* **1647**:18–23 DOI 10.1016/S1570-9639(03)00042-6.

Banning NC, Gleeson DB, Grigg AH, Grant CD, Andersen GL, Brodie EL, Murphy DV. 2011. Soil microbial community successional patterns during forest ecosystem restoration. *Applied and Environmental Microbiology* **77**:6158–6164 DOI 10.1128/AEM.00764-11.

Beck DAC, McTaggart TL, Setboonsarng U, Vorobev A, Kalyuzhnaya MG, Ivanova N, Goodwin L, Woyke T, Lidstrom ME, Chistoserdova L. 2014. The expanded diversity of Methylophilaceae from Lake Washington through cultivation and genomic sequencing of novel ecotypes. *PLoS ONE* **9**:e102458 DOI 10.1371/journal.pone.0102458.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**:289–300.

Berendse F. 1994. Litter decomposability—a neglected component of plant fitness. *Journal of Ecology* **82**:187–190 DOI 10.2307/2261398.

Biddle JF, Lipp JS, Lever MA, Lloyd KG, Sørensen KB, Anderson R, Fredricks HF, Elvert M, Kelly TJ, Schrag DP, Sogin ML, Brenchley JE, Teske A, House CH, Hinrichs K-U. 2006. Heterotrophic Archaea dominate sedimentary subsurface ecosystems off Peru. *Proceedings of the National Academy of Sciences of the United States of America* **103**:3846–3851 DOI 10.1073/pnas.0600035103.

Blazewicz SJ, Schwartz E, Firestone MK. 2014. Growth and death of bacteria and fungi underlie rainfall-induced carbon dioxide pulses from seasonally dried soil. *Ecology* **95**:1162–1172 DOI 10.1890/13-1031.1.

Bogart JA, Lewis AJ, Schelter EJ. 2015. DFT study of the active site of the XoxF-type natural, cerium-dependent methanol dehydrogenase enzyme. *Chemistry* **21**:1743–1748 DOI 10.1002/chem.201405159.

Brooks B, Mueller RS, Young JC, Morowitz MJ, Hettich RL, Banfield JF. 2015. Strain-resolved microbial community proteomics reveals simultaneous aerobic and anaerobic function during gastrointestinal tract colonization of a preterm infant. *Frontiers in Microbiology* **6**: Article 654 DOI 10.3389/fmicb.2015.00654.

Castelle CJ, Wrighton KC, Thomas BC, Hug LA, Brown CT, Wilkins MJ, Frischkorn KR, Tringe SG, Singh A, Markillie LM, Taylor RC, Williams KH, Banfield JF. 2015. Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Current Biology* **25**:690–701 DOI 10.1016/j.cub.2015.01.014.

Chistoserdova L. 2011. Modularity of methylotrophy, revisited. *Environmental Microbiology* **13**:2603–2622 DOI 10.1111/j.1462-2920.2011.02464.x.

Corpe WA, Rheem S. 1989. Ecology of the methylotrophic bacteria on living leaf surfaces. *FEMS Microbiology Letters* **62**:243–249 DOI 10.1016/0378-1097(89)90248-6.

Costello AM, Lidstrom ME. 1999. Molecular characterization of functional and phylogenetic genes from natural populations of methanotrophs in lake sediments. *Applied and Environmental Microbiology* **65**:5066–5074.

Cruz-Martínez K, Suttle KB, Brodie EL, Power ME, Andersen GL, Banfield JF. 2009. Despite strong seasonal responses, soil microbial consortia are more resilient to long-term changes in rainfall than overlying grassland. *ISME Journal* **3**:738–744 DOI 10.1038/ismej.2009.16.

Delmont TO, Eren AM, Maccario L, Prestat E, Esen OC, Pelletier E, Le Paslier D, Simonet P, Vogel TM. 2015. Reconstructing rare soil microbial genomes using *in situ* enrichments and metagenomics. *Frontiers in Microbiology* **6**: Article 358 DOI 10.3389/fmicb.2015.00358.

Delmotte N, Knief C, Chaffron S, Innerebner G, Roschitzki B, Schlapbach R, Von Mering C, Vorholt JA. 2009. Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **106**:16428–16433 DOI 10.1073/pnas.0905240106.

Dixon P. 2003. VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science* **14**:927–930.

Edgar R. 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**:1–19 DOI 10.1186/1471-2105-5-113.

Edgar RC. 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**:1792–1797 DOI 10.1093/nar/gkh340.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**:2460–2461 DOI 10.1093/bioinformatics/btq461.

Elias JE, Gygi SP. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* **4**:207–214 DOI 10.1038/nmeth1019.

Evans PN, Parks DH, Chadwick GL, Robbins SJ, Orphan VJ, Golding SD, Tyson GW. 2015. Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science* **350**:434–438 DOI 10.1126/science.aac7745.

Evans SE, Wallenstein MD. 2012. Soil microbial community response to drying and rewetting stress: does historical precipitation regime matter? *Biogeochemistry* **109**:101–116 DOI 10.1007/s10533-011-9638-3.

Eyice O, Namura M, Chen Y, Mead A, Samavedam S, Schafer H. 2015. SIP metagenomics identifies uncultivated Methylophilaceae as dimethylsulphide degrading bacteria in soil and lake sediment. *ISME Journal* **9**:2336–2348 DOI 10.1038/ismej.2015.37.

Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert JA, Wall DH, Caporaso JG. 2012. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences of the United States of America* **109**:21390–21395 DOI 10.1073/pnas.1215210110.

Fischer H, Ingwersen J, Kuzyakov Y. 2010. Microbial uptake of low-molecular-weight organic substances out-competes sorption in soil. *European Journal of Soil Science* **61**:504–513 DOI 10.1111/j.1365-2389.2010.01244.x.

Gagen EJ, Huber H, Meador T, Hinrichs KU, Thomm M. 2013. Novel cultivation-based approach to understanding the Miscellaneous Crenarchaeotic Group (MCG) archaea from sedimentary ecosystems. *Applied and Environmental Microbiology* **79**:6400–6406 DOI 10.1128/AEM.02153-13.

Goldfarb KC, Karaoz U, Hanson CA, Santee CA, Bradford MA, Treseder KK, Wallenstein MD, Brodie EL. 2011. Differential growth responses of soil bacterial taxa to carbon substrates of varying chemical recalcitrance. *Frontiers in Microbiology* **2**: Article 94 DOI 10.3389/fmicb.2011.00094.

He J, Xu Z, Hughes J. 2006. Molecular bacterial diversity of a forest soil under residue management regimes in subtropical Australia. *FEMS Microbiology Ecology* **55**:38–47 DOI 10.1111/j.1574-6941.2005.00006.x.

Henckel T, Friedrich M, Conrad R. 1999. Molecular analyses of the methane-oxidizing microbial community in rice field soil by targeting the genes of the 16S rRNA, particulate methane monooxygenase, and methanol dehydrogenase. *Applied and Environmental Microbiology* **65**:1980–1990.

Herzberger AJ, Duncan DS, Jackson RD. 2014. Bouncing back: plant-associated soil microbes respond rapidly to prairie establishment. *PLoS ONE* **9**:e115775 DOI 10.1371/journal.pone.0115775.

Hofmann K, Pauli H, Praeg N, Wagner AO, Illmer P. 2016. Methane-cycling microorganisms in soils of a high-alpine altitudinal gradient. *FEMS Microbiology Ecology* **92(3)**: fiw009 DOI 10.1093/femsec/fiw009.

Hug LA, Thomas BC, Brown CT, Frischkorn KR, Williams KH, Tringe SG, Banfield JF. 2015. Aquifer environment selects for microbial species cohorts in sediment and groundwater. *ISME Journal* **9**:1846–1856 DOI 10.1038/ismej.2015.2.

Hug LA, Thomas BC, Sharon I, Brown CT, Sharma R, Hettich RL, Wilkins MJ, Williams KH, Singh A, Banfield JF. 2016. Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. *Environmental Microbiology* **18**:159–173 DOI 10.1111/1462-2920.12930.

Hultman J, Waldrop MP, Mackelprang R, David MM, McFarland J, Blazewicz SJ, Harden J, Turetsky MR, McGuire AD, Shah MB, VerBerkmoes NC, Lee LH, Mavrommatis K, Jansson JK. 2015. Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature* **521**:208–212 DOI 10.1038/nature14238.

Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**:119 DOI 10.1186/1471-2105-11-119.

Hyatt D, Pan C. 2012. Exhaustive database searching for amino acid mutations in proteomes. *Bioinformatics* **28**:1895–1901 DOI 10.1093/bioinformatics/bts274.

Jones DL, Shannon D, Murphy DV, Farrar J. 2004. Role of dissolved organic nitrogen (DON) in soil N cycling in grassland soils. *Soil Biology and Biochemistry* **36**:749–756 DOI 10.1016/j.soilbio.2004.01.003.

Kandeler E, Mosier AR, Morgan JA, Milchunas DG, King JY, Rudolph S, Tscherko D. 2006. Response of soil microbial biomass and enzyme activities to the transient elevation of carbon dioxide in a semi-arid grassland. *Soil Biology and Biochemistry* **38**:2448–2460 DOI 10.1016/j.soilbio.2006.02.021.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research* **40**:D109–D114 DOI 10.1093/nar/gkr988.

Karl DM, Beversdorf L, Bjorkman KM, Church MJ, Martinez A, Delong EF. 2008. Aerobic production of methane in the sea. *Nature Geoscience* **1**:473–478 DOI 10.1038/ngeo234.

Keltjens JT, Pol A, Reimann J, Op Den Camp HJM. 2014. PQQ-dependent methanol dehydrogenases: rare-earth elements make a difference. *Applied Microbiology and Biotechnology* **98**:6163–6183 DOI 10.1007/s00253-014-5766-8.

Kolb S. 2009. Aerobic methanol-oxidizing Bacteria in soil. *FEMS Microbiology Letters* **300**:1–10 DOI 10.1111/j.1574-6968.2009.01681.x.

Kubo K, Lloyd KG, Biddle JF, Amann R, Teske A, Knittel K. 2012. Archaea of the Miscellaneous Crenarchaeotal Group are abundant, diverse and widespread in marine sediments. *ISME Journal* **6**:1949–1965 DOI 10.1038/ismej.2012.37.

Kuramae EE, Yergeau E, Wong LC, Pijl AS, Van Veen JA, Kowalchuk GA. 2012. Soil characteristics more strongly influence soil bacterial communities than land-use type. *FEMS Microbiology Ecology* **79**:12–24 DOI 10.1111/j.1574-6941.2011.01192.x.

Lauber CL, Hamady M, Knight R, Fierer N. 2009. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Applied and Environmental Microbiology* **75**:5111–5120 DOI 10.1128/AEM.00335-09.

Li Z, Wang Y, Yao Q, Justice NB, Ahn TH, Xu D, Hettich RL, Banfield JF, Pan C. 2014. Diverse and divergent protein post-translational modifications in two growth stages of a natural microbial community. *Nature Communications* **5**: Article 4405 DOI 10.1038/ncomms5405.

Lloyd KG, Schreiber L, Petersen DG, Kjeldsen KU, Lever MA, Steen AD, Stepanauskas R, Richter M, Kleindienst S, Lenk S, Schramm A, Jorgensen BB. 2013. Predominant archaea in marine sediments degrade detrital proteins. *Nature* **496**:215–218 DOI 10.1038/nature12033.

Luo C, Rodriguez-R LM, Johnston ER, Wu L, Cheng L, Xue K, Tu Q, Deng Y, He Z, Shi JZ, Yuan MM, Sherry RA, Li D, Luo Y, Schuur EAG, Chain P, Tiedje JM, Zhou J, Konstantinidis KT. 2014. Soil microbial community responses to a decade of warming as revealed by comparative metagenomics. *Applied and Environmental Microbiology* **80**:1777–1786 DOI 10.1128/AEM.03712-13.

Markham P, Robson GD, Bainbridge BW, Trinci AP. 1993. Choline: its role in the growth of filamentous fungi and the regulation of mycelial morphology. *FEMS Microbiology Reviews* **10**:287–300.

Mau RL, Liu CM, Aziz M, Schwartz E, Dijkstra P, Marks JC, Price LB, Keim P, Hungate BA. 2015. Linking soil bacterial biodiversity and soil carbon stability. *ISME Journal* **9**:1477–1480 DOI 10.1038/ismej.2014.205.

McKissock I, Gilkes RJ, Walker EL. 2002. The reduction of water repellency by added clay is influenced by clay and soil properties. *Applied Clay Science* **20**:225–241 DOI 10.1016/S0169-1317(01)00074-6.

Metcalf WW, Griffin BM, Cicchillo RM, Gao J, Janga SC, Cooke HA, Circello BT, Evans BS, Martens-Habbena W, Stahl DA, Van der Donk WA. 2012. Synthesis of methylphosphonic acid by marine microbes: a source for methane in the aerobic ocean. *Science* **337**:1104–1107 DOI 10.1126/science.1219875.

Mosier AC, Li Z, Thomas BC, Hettich RL, Pan C, Banfield JF. 2015. Elevated temperature alters proteomic responses of individual organisms within a biofilm community. *ISME Journal* **9**:180–194 DOI 10.1038/ismej.2014.113.

Muyzer G, De Waal EC, Uitterlinden AG. 1993. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Applied and Environmental Microbiology* **59**:695–700.

Nawrocki EP, Kolbe DL, Eddy SR. 2009. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**:1335–1337 DOI 10.1093/bioinformatics/btp157.

Nercessian O, Noyes E, Kalyuzhnaya MG, Lidstrom ME, Chistoserdova L. 2005. Bacterial populations active in metabolism of C1 compounds in the sediment of Lake Washington, a freshwater lake. *Applied and Environmental Microbiology* **71**:6885–6899 DOI 10.1128/AEM.71.11.6885-6899.2005.

Nesvizhskii AI, Aebersold R. 2005. Interpretation of shotgun proteomic data: the protein inference problem. *Molecular & Cellular Proteomics* **4**:1419–1440 DOI 10.1074/mcp.R500012-MCP200.

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. 1999. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **27**:29–34 DOI 10.1093/nar/27.1.29.

Op den Camp HJM, Islam T, Stott MB, Harhangi HR, Hynes A, Schouten S, Jetten MSM, Birkeland NK, Pol A, Dunfield PF. 2009. Environmental, genomic and taxonomic perspectives on methanotrophic Verrucomicrobia. *Environmental Microbiology Reports* **1**:293–306 DOI 10.1111/j.1758-2229.2009.00022.x.

Osborn AM, Moore ERB, Timmis KN. 2000. An evaluation of terminal-restriction fragment length polymorphism (T-RFLP) analysis for the study of microbial community structure and dynamics. *Environmental Microbiology* **2**:39–50 DOI 10.1046/j.1462-2920.2000.00081.x.

Pan C, Banfield JF. 2014. Quantitative metaproteomics: functional insights into microbial communities. *Methods in Molecular Biology* **1096**:231–240 DOI 10.1007/978-1-62703-712-9_18.

Pell J, Hintze A, Canino-Koning R, Howe A, Tiedje JM, Brown CT. 2012. Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proceedings of the National Academy of Sciences of the United States of America* **109**:13272–13277 DOI 10.1073/pnas.1121464109.

Peltoniemi K, Laiho R, Juottonen H, Kiikkilä O, Mäkiranta P, Minkkinen K, Pennanen T, Penttilä T, Sarjala T, Tuittila E-S, Tuomivirta T, Fritze H. 2015. Microbial

ecology in a future climate: effects of temperature and moisture on microbial communities of two boreal fens. *FEMS Microbiology Ecology* **91(7)**: fiv062 DOI 10.1093/femsec/fiv062.

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 DOI 10.1093/bioinformatics/bts174.

Pester M, Rattei T, Flechl S, Gröngröft A, Richter A, Overmann J, Reinhold-Hurek B, Loy A, Wagner M. 2012. AmoA-based consensus phylogeny of ammonia-oxidizing archaea and deep sequencing of amoA genes from soils of four different geographic regions. *Environmental Microbiology* **14**:525–539 DOI 10.1111/j.1462-2920.2011.02666.x.

Piper CL, Siciliano SD, Winsley T, Lamb EG. 2015. Smooth brome invasion increases rare soil bacterial species prevalence, bacterial species richness and evenness. *Journal of Ecology* **103**:386–396 DOI 10.1111/1365-2745.12356.

Placella SA, Brodie EL, Firestone MK. 2012. Rainfall-induced carbon dioxide pulses result from sequential resuscitation of phylogenetically clustered microbial groups. *Proceedings of the National Academy of Sciences of the United States of America* **109**:10931–10936 DOI 10.1073/pnas.1204306109.

Pol A, Barends TRM, Dietl A, Khadem AF, Eygensteyn J, Jetten MSM, Op den Camp HJM. 2014. Rare earth metals are essential for methanotrophic life in volcanic mudpots. *Environmental Microbiology* **16**:255–264 DOI 10.1111/1462-2920.12249.

Prober SM, Leff JW, Bates ST, Borer ET, Firn J, Harpole WS, Lind EM, Seabloom EW, Adler PB, Bakker JD, Cleland EE, Decrappeo NM, Delorenze E, Hagenah N, Hautier Y, Hofmockel KS, Kirkman KP, Knops JMH, La Pierre KJ, Macdougall AS, McCulley RL, Mitchell CE, Risch AC, Schuetz M, Stevens CJ, Williams RJ, Fierer N. 2015. Plant diversity predicts beta but not alpha diversity of soil microbes across grasslands worldwide. *Ecology Letters* **18**:85–95 DOI 10.1111/ele.12381.

Probst AJ, Castelle CJ, Singh A, Brown CT, Anantharaman K, Sharon I, Hug LA, Burstein D, Emerson JB, Thomas BC, Banfield JF. 2016. Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO<inf>2</inf> concentrations. *Environmental Microbiology* Epub ahead of print Apr 26 2016 DOI 10.1111/1462-2920.13362.

Pruesse E, Peplies J, Glöckner FO. 2012. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**:1823–1829 DOI 10.1093/bioinformatics/bts252.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO. 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* **35**:7188–7196 DOI 10.1093/nar/gkm864.

Radajewski S, Webster G, Reay DS, Morris SA, Ineson P, Nedwell DB, Prosser JI, Murrell JC. 2002. Identification of active methylotroph populations in an acidic forest soil by stable-isotope probing. *Microbiology* **148**:2331–2342.

R Core Team. 2015. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. *Available at http://www.R-project.org/*.

Reinsch S, Michelsen A, Sárossy Z, Egsgaard H, Schmidt IK, Jakobsen I, Ambus P. 2014. Short-term utilization of carbon by the soil microbial community under future climatic conditions in a temperate heathland. *Soil Biology and Biochemistry* **68**:9–19 DOI 10.1016/j.soilbio.2013.09.014.

Schimel JP, Schaeffer SM. 2012. Microbial control over carbon cycling in soil. *Frontiers in Microbiology* **3**: Article 348 DOI 10.3389/fmicb.2012.00348.

Scott JW, Rasche ME. 2002. Purification, overproduction, and partial characterization of β-RFAP synthase, a key enzyme in the methanopterin biosynthesis pathway. *Journal of Bacteriology* **184**:4442–4448 DOI 10.1128/jb.184.16.4442-4448.2002.

Sharon I, Kertesz M, Hug LA, Pushkarev D, Blauwkamp TA, Castelle CJ, Amirebrahimi M, Thomas BC, Burstein D, Tringe SG, Williams KH, Banfield JF. 2015. Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. *Genome Research* **25**:534–543 DOI 10.1101/gr.183012.114.

Skovran E, Gomez NCM. 2015. Just add lanthanides. *Science* **348**:862–863 DOI 10.1126/science.aaa9091.

Stacheter A, Noll M, Lee CK, Selzer M, Glowik B, Ebertsch L, Mertel R, Schulz D, Lampert N, Drake HL, Kolb S. 2013. Methanol oxidation by temperate soils and environmental determinants of associated methylotrophs. *ISME Journal* **7**:1051–1064 DOI 10.1038/ismej.2012.167.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313 DOI 10.1093/bioinformatics/btu033.

Suttle KB, Thomsen MA, Power ME. 2007. Species interactions reverse grassland responses to changing climate. *Science* **315**:640–642 DOI 10.1126/science.1136401.

Sutton R, Sposito G. 2005. Molecular structure in soil humic substances: the new view. *Environmental Science and Technology* **39**:9009–9015 DOI 10.1021/es050778q.

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**:1282–1288 DOI 10.1093/bioinformatics/btm098.

Swenson TL, Jenkins S, Bowen BP, Northen TR. 2015. Untargeted soil metabolomics methods for analysis of extractable organic matter. *Soil Biology and Biochemistry* **80**:189–198 DOI 10.1016/j.soilbio.2014.10.007.

Taubert M, Grob C, Howat AM, Burns OJ, Dixon JL, Chen Y, Murrell JC. 2015. XoxF encoding an alternative methanol dehydrogenase is widespread in coastal marine environments. *Environmental Microbiology* **17**:3937–3948 DOI 10.1111/1462-2920.12896.

Verastegui Y, Cheng J, Engel K, Kolczynski D, Mortimer S, Lavigne J, Montalibet J, Romantsov T, Hall M, McConkey BJ, Rose DR, Tomashek JJ, Scott BR, Charles TC, Neufeld JD. 2014. Multisubstrate isotope labeling and metagenomic analysis of active soil bacterial communities. *mBio* **5**:e44203 DOI 10.1128/mBio.01157-14.

Veresoglou SD, Thornton B, Menexes G, Mamolos AP, Veresoglou DS. 2012. Soil fertilization leads to a decline in between-samples variability of microbial community δ13C profiles in a grassland fertilization experiment. *PLoS ONE* **7**:e44203 DOI 10.1371/journal.pone.0044203.

Vizcaíno JA, Csordas A, Del-Toro N, Dianes JA, Griss J, Lavidas I, Mayer G, Perez-Riverol Y, Reisinger F, Ternent T, Xu QW, Wang R, Hermjakob H. 2016. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Research* **44**:D447–D456 DOI 10.1093/nar/gkv1145.

Wang Y, Ahn TH, Li Z, Pan C. 2013. Sipros/ProRata: a versatile informatics system for quantitative community proteomics. *Bioinformatics* **29**:2064–2065 DOI 10.1093/bioinformatics/btt329.

Washburn MP, Wolters D, Yates 3rd JR. 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology* **19**:242–247 DOI 10.1038/85686.

Wedin DA, Tilman D. 1990. Species effects on nitrogen cycling: a test with perennial grasses. *Oecologia* **84**:433–441 DOI 10.1007/BF00328157 .

White RA, Bottos EM, Roy Chowdhury T, Zucker JD, Brislawn CJ, Nicora CD, Fansler SJ, Glaesemann KR, Glass K, Jansson JK. 2016. Moleculo long-read sequencing facilitates assembly and genomic binning from complex soil metagenomes. *mSystems* **1**(**3**):e00045-16 DOI 10.1128/mSystems.00045-16.

White R, Murray A, Rohweder M. 2000. *Pilot analysis of global ecosystems: grassland ecosystems*. Washington, D.C.: World Resources Institute.

Wisniewski JR, Zougman A, Nagaraj N, Mann M. 2009. Universal sample preparation method for proteome analysis. *Nature Methods* **6**:359–362 DOI 10.1038/nmeth.1322.

Mediterranean grassland soil C-N compound turnover is mediated by genomically divergent organisms, depth stratified, and rainfall dependent

Spencer Diamond[1], Peter F. Andeer[2], Zhou Li[3], Alexander Crits-Christoph[4], David Burstein[1,#], Karthik Anantharaman[1,+], Katherine R. Lane[1], Brian C. Thomas[1], Chongle Pan[3,†], Trent R. Northen[2,6] and Jillian F. Banfield[1,5,*]

[1]Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, CA, USA
[2]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, Berkeley, CA, 94720, USA.
[3]Oak Ridge National Laboratory, Oak Ridge, TN, USA
[4]Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, CA, USA
[5]Department of Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, CA, USA
[6]Joint Genome Institute, Lawrence Berkeley National Laboratory, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA

[#]Current Address: School of Molecular Cell Biology and Biotechnology, Tel Aviv University, Tel Aviv, Israel

[+]Current Address: Department of Bacteriology, University of Wisconsin, Madison, WI, USA

[†]Current Address: School of Computer Science and Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK, USA

***Corresponding author:**
Jillian F. Banfield,
Energy Biosciences Building
2151 Berkeley Way
Berkeley, CA 94720-5230
510-643-2155
jbanfield@berkeley.edu

Abstract

Soil microbial activity drives the carbon and nitrogen cycles and is an important determinant of atmospheric trace gas turnover, yet most soils are dominated by microorganisms with unknown metabolic capacities. Even Acidobacteria, among the most abundant bacteria in soil, remain poorly characterized, and functions across groups such as Verrucomicrobia, Gemmatimonadetes, Chloroflexi, and Rokubacteria are understudied. Here, we resolved sixty metagenomic, and twenty proteomic datasets from a mediterranean grassland soil ecosystem and recovered 793 near-complete microbial genomes from 18 phyla, representing around one third of all microorganisms detected. Importantly, this enabled extensive genomics-based metabolic predictions for these communities. Acidobacteria from multiple previously unstudied classes have genomes that encode large enzyme complements for complex carbohydrate degradation. Alternatively, most microorganisms encode carbohydrate esterases that strip readily accessible methyl and acetyl groups from polymers like pectin and xylan, forming methanol and acetate, the availability of which could explain high prevalence of C1 metabolism and acetate utilization in genomes. Microorganism abundances among samples collected at three soil depths and under natural and amended rainfall regimes indicate statistically higher associations of inorganic nitrogen metabolism and carbon degradation in deep and shallow soils, respectively. This partitioning decreased in samples under extended spring rainfall indicating long term climate alteration can affect both carbon and nitrogen cycling. Overall, by leveraging natural and experimental gradients with genome-resolved metabolic profiles, we link microorganisms lacking prior genomic characterization to specific roles in complex carbon, C1, nitrate, and ammonia transformations and constrain factors that impact their distributions in soil.

Introduction

Grassland ecosystems cover 26% of all land area, store 34% of global terrestrial carbon, and comprise 80% of agriculturally productive land[1,2]. Therefore grasslands have a significant impact on global soil carbon storage, trace gas emissions, and economic productivity[1,2]. Identifying microorganism capacities for carbon and nitrogen turnover is critical, as microorganisms ultimately determine how grassland soils cycle carbon and nitrogen, and emit or absorb trace gases[3,4] (In the context of this manuscript microorganisms only refers to Bacteria and Archaea).

One of the biggest challenges in studying the metabolism of soil microbial communities is that most of the microorganisms have only been detected using 16S rRNA surveys[5,6]. While studies have been undertaken to link amplified metabolic genes or 16S rRNA gene abundances with soil trace gas fluxes or environmental conditions[7-11], the large number of soil-associated microorganisms not represented by genomes precludes meaningful predictions of relationships between microorganism types and their biogeochemical functions.

The metabolic capacities of soil-associated microorganisms can be investigated if genomes can be reconstructed from soil samples[12-14]. However, this is notoriously difficult, as most soils have extremely high microbial diversity[15]. To date, few soil datasets have been even partially genomically resolved[13,16], but recently it was shown that broad genomic resolution and community metabolic functions could be deduced in metagenomic studies targeting permafrost[12].

Here, we applied deep metagenomic sequencing and metaproteomic analyses to sub-root zone samples from a grassland soil ecosystem from a mediterranean climate. Mediterranean grassland soils are of particular interest as they have not been genomically characterized, and undergo strong seasonal drying and re-wetting that uniquely structures their microbial communities[17,18]. A subset of the soils in this study are currently undergoing a rainfall

extension climate change experiment[19]. Despite the presence of thousands of species at low abundance levels and strain heterogeneity, we successfully reconstructed non-redundant draft-quality genomes that account for the majority of microorganisms detected by abundance. Overall our data reveal important carbon and nitrogen turnover functions in understudied microbial groups, show a stark metabolic and phylogenetic stratification across soil depths, and support climate change as a factor that can significantly alter the carbon and nitrogen turnover capacity of soil microbial communities.

Results

Soil sampling and assembly. We collected 60 soil samples from 10-20 cm (just below the root zone), 20-30 cm, and 30-40 cm from a grassland meadow within the Angelo Coastal Range Reserve in Northern California (Supplementary Fig. 1). Three of the six sampling sites had been subjected to over 14 years of rainfall amendment to simulate a predicted climate change scenario for northern California[19]. In total, we generated 1.2 Tb of raw read data which assembled into 67 Gbp of contiguous sequence. Of this, 47 Gbp (70.2%) of the assembled sequences were >1 kb in length. On average 36.4% of reads mapped back assemblies, and for some samples this mapping was as high as 64.7% (Supplementary Table 1).

A species richness census reveals extensive sampling of soil microbial diversity. Although our approach overall is genome-centric, many microorganisms were at too low abundance to be represented by draft genomes. Thus, we used ribosomal protein S3 (rpS3) to conduct a census of the microbial diversity found at the site and to quantify relative organism abundances[20]. Across our 60 metagenomic assemblies we identified 10,158 rpS3 sequences (169±93 per sample), which were grouped into 3325 non-redundant clusters (Methods) that

74 approximate species groups (SGs) (Supplementary Table 2, Supplementary Fig. 2, and
75 Supplementary Data 1-2).
76 Using our rpS3 sequences as phylogenetic markers we initially classified all of the
77 microorganisms detected at the phylum and class levels. We detected 26 distinct phylum-level
78 lineages, and the topology of the rpS3 tree suggested that most phyla are represented by few
79 class level groups with high degrees of genus and species heterogeneity. We also found that
80 the abundances of closely related microorganisms could be highly variable, differing in
81 abundance by a factor of 10 (Supplementary Fig. 3 and Supplementary Data 3).
82 In agreement with many previous soil surveys[5,6], we found that Verrucomicobia and
83 Acidobacteria were the most relatively abundant lineages across our site (Fig. 1a). Generally,
84 coverage was disproportionately concentrated in a small subset of SGs, and approximately 13%
85 (443) of the detected microorganisms accounted for 50% of the total read coverage (Fig. 1c).
86 Some microorganisms, such as specific Nitrospirae and Euryarchaeota, had high relative
87 abundance despite their phylum as a whole exhibiting low relative abundance (Fig. 1a and 1c).
88 Thus, while some phyla do not collectively account for a highly fraction of the reconstructed
89 microbiomes, individual microorganisms belonging to these phyla may be highly abundant.
90
91 **Spatial variation and treatment, but not time of sampling, contribute significant variance**
92 **to microorganism abundance.** To visualize the influence of depth, sampling location, sampling
93 date, and rainfall amendment on the abundance of SGs, we applied non-metric
94 multidimensional scaling (NMDS) ordination to the weighted Unifrac distance matrix of SG
95 coverage (Fig. 1b, Supplementary Table 3-4, Supplementary Data 4). Subsequently we used
96 the Multi-Response Permutation Procedure (MRPP) to test the significance and strength of
97 each variable's influence. The results indicate that sampling depth, sampling location, and
98 rainfall amendment had significant effects on relative microorganism abundance and

composition across samples (Fig. 1b). Sampling depth was the most influential factor (C = 0.26; p = $1e^{-4}$), followed by sampling location (C = 0.12; p = $2e^{-4}$), and rainfall amendment (C = 0.02; p = 0.04). While rainfall amendment showed a consistent effect, its effect occurs relative to sampling location (Fig. 1b). A large sample number was critical in observing this relationship, and was important for isolating the weaker effect caused by rainfall extension. Alternatively, we found that the date a sample was collected did not significantly influence overall SG variability, despite samples being collected over a 31-day period covering the transition from the dry to rainy season (Fig. 1b and Supplementary Fig. 1).

A hybrid binning method resolved genomes from previously unsequenced lineages. Genomes reconstructed from each sample were used to link metabolic functions to specific microorganisms (Methods). We recovered 10,463 genomic bins with an average of 174±87 binned genomes per sample. After clustering bins based on the SGs assigned to their rpS3 gene, and filtering for estimated completeness > 70% and contamination < 10%, we recovered 793 unique microbial genomes (Supplementary Table 5).

Our reconstructed genomes represent 24% of SGs by number, however these genomes represent more than half (53%) of the SGs by total coverage (Fig. 1a). 204 genomes were from microorganisms in the lowest quartile of total abundance (Fig. 1c). Importantly, we recovered 115 high quality genomes (> 95% estimated completeness) across 15 of the 26 microbial phyla detected at the site (Supplementary Table 5).

A more detailed phylogenetic analysis using both a concatenated set of 15 ribosomal proteins (rp15) and 16S rRNA sequences indicated that we have significantly expanded the genomic coverage across a number of poorly sequenced soil lineages (Fig. 2, Supplementary Fig. 4-5, Supplementary Table 5-6, and Supplementary Data 5-8). Many genomes from

unsequenced lineages were relatively abundant microorganisms at our site. In particular, we recovered 145 near complete Acidobacterial genomes from 15 class-level lineages, four of which have no previously sequenced representative (Gp18, Gp5, Gp11, and Gp2) (Fig. 2 and Supplementary Fig. 4-5)**.** We also found phylogenetic overlap between our Acidobacterial genomes and previously recovered but unclassified Acidobacterial genomes from a subsurface aquifer sediment in Rifle, Colorado[21]. By including genomes from both the Rifle and Angelo sites in our phylogenetic tree we were able to assign 17 genomes to Acidobacterial Classes Gp7, Gp22, and Gp17 for which there was no previous class-level genomic information (Supplementary Fig. 5).

The majority of our Chloroflexi genomes came from four unsequenced or poorly sequenced class-level lineages. Nine genomes affiliate with a group referred to as CHLX from Rifle aquifer sediment[21], and 32 genomes phylogenetically place with a second lineage that includes one genome from Rifle sediment and one from arctic soil[13]. We also recovered 96 genomes from two class-level lineages within Chloroflexi with no previously sequenced representatives, hereafter referred to as ANG-CHLX1 and ANG-CHLX2 (Fig. 2, Supplementary Fig. 4). The ANG-CHLX1 and ANG-CHLX2 clades form a strongly supported group basal to RIF-CHLX genomes and all known Chloroflexi lineages.

The soil proteome indicates a high prevalence of C1, pentose sugar, and small molecule metabolism. We used shotgun proteomic data from 20 samples to provide insight into abundant functions in situ (Methods), and to guide or metabolic analysis of the reconstructed genomes. Overall, we identified 55,665 proteins with at least one uniquely mapped peptide that was detected with high mass accuracy. In total, 60% of the proteins identified could be assigned

to one of 393 functional orthology groups (Supplementary Tables 7-8 and Supplementary Data 9).

The most abundant proteins identified were ABC transporters for sugars and amino acids, pentose sugar processing enzymes, and enzymes degrading small C1 and nitrogen containing compounds including formamidase, carbon monoxide dehydrogenase, and methanol dehydrogenase (Supplementary Fig. 6 and Supplementary Results). A high abundance of xoxF-type methanol dehydrogenases had been previously reported from proteomics at this site[14]. In this study, we also detected high abundances of proteins annotated as carbon monoxide dehydrogenases (coxL), including coxL-TypeI that functions in the oxidation of CO, and others. Genomic studies have indicated widespread distribution of diverse coxL subtypes in soils[9,22,23], suggesting that subtypes other than TypeI may be important and overlooked small molecule dehydrogenases with unknown specificity.

Genome metabolic profiling identifies prevalent metabolism of small molecules and nitrogen cycling processes in unexpected microorganisms. Given the prevalence of enzymes that turnover low molecular weight compounds we targeted their genes in our analyses of genome metabolic potential. The dbCAN and KEGG databases were used to profile reconstructed genomes[24,25] (See Methods, Supplementary Figs. 7-9, Supplementary Tables 9-13, and Supplementary Data 10-14).

Methanol dehydrogenases were detected in 187 genomes, and all methanol dehydrogenases identified were of the XoxF-type (Fig. 3a, Supplementary Fig. 7). These genes were abundant in Gemmatimonadetes and Rokubacteria, but also were detected in Gp1, Gp5, and Gp6 Acidobacterial genomes and 4 phyla of Proteobacteria (Fig. 3a). 90 genomes encode formamidase (amiF), including 26 Chloroflexi and 30 Rokubacteria (Fig. 3a). Formamidase contributes to both formate and ammonia pools via the breakdown of formamide that may

171 originate from amino acid catabolism[25]. Using coxL as a marker for coxLMS type CO-
172 dehydrogenases[9] we detected 1889 coxL homologues encoded in 466 genomes. However, only
173 coxL-TypeI is known to metabolize CO[9,22]. We note that coxL-TypeI genes were encoded in 59
174 Chloroflexi genomes, with the majority being from ANG-CHLX1 and ANG-CHLX2 clades (Fig.
175 3a). However, the vast majority of coxL proteins were subtypes other than coxL-TypeI
176 (Supplementary Fig. 8).
177 Bacteroidetes and Acidobacteria genomes encode the largest number of Carbohydrate
178 Active Enzymes (CAZy enzymes), but Acidobacteria far exceed Bacteroidetes in terms of both
179 total genomes detected (152 vs 5) and relative abundance (16% vs 0.2% across all
180 communities) (Fig. 2, 3c, Supplementary Fig. 4). Acidobacteria also have the highest diversity of
181 CAZy enzyme types, with 73% of CAZy families detected in at least one member of this phylum
182 (Fig. 3c). Acidobacterial genomes from classes Gp1 and Gp3 are known to contain large
183 numbers of CAZy enzyme genes[26]. Here we identified 9 Acidobacterial classes containing
184 genomes that encode >100 CAZy enzymes, including the previously unsequenced classes Gp2,
185 Gp11, and Gp18 (Supplementary Fig. 10). This significantly expands the metabolic potential for
186 complex carbohydrate turnover across the Acidobacteria phylum.
187 Across CAZy enzymes we noted a particularly high proportion of carbohydrate esterases
188 (22%; Fig. 3a, 3c, and Supplementary Fig. 10). Types CE1 and CE4 account for 56% of all
189 carbohydrate esterases identified, and liberate acetate from a broad spectrum of complex plant
190 and microbial polymers[27,28]. Of the 793 genomes analyzed 81% contain either CE1 or CE4 as
191 well as an encoded acetyl-coA synthetase to incorporate liberated acetate (Supplementary Fig.
192 10)
193 In analyzing the genomic capacity to mediate inorganic nitrogen transformations we
194 found most microorganisms only encode a single transformation reaction, and that nitrite is the
195 most common reaction substrate (Fig. 3a, 3b, and Supplementary Table 10). We did not detect

any genome with the potential for complete denitrification, or complete nitrification via ammonia oxidation (Fig. 3b). Also, we found only two genomes classified as Bacteroidetes encoding the enzyme nosZ, which may indicate limited N_2O turnover potential in this system (Fig. 3b).

Of the 49 genomes encoding nirK, 12 were Gemmatimonadetes, a genomically under sampled phylum that is not normally linked to nitrite conversion to nitric oxide (Fig. 3b). Many Gemmatimonadetes with nirK were also relatively abundant (Supplementary Table 10). The gene norB, which converts nitric oxide (NO) to N_2O, was exclusively found within genomes of Acidobacteria (Fig. 3b). While five acidobacterial classes had been previously reported to encode norB[29], we additionally detected these genes in Acidobacteria from Gp4, Gp5, and Gp13 suggesting a widespread capacity for nitric oxide reduction across the acidobacterial phylum (Supplementary Table 10).

Microorganisms are phylogenetically and functionally stratified by depth. 391 genomes significantly increased and 179 decreased in abundance with increasing soil depth. Thus, the majority of assembled genomes (72%) exhibit abundance patterns stratified by depth (Fig. 2 and Supplementary Table 5). All Archaeal lineages as well as Rokubacteria and Gemmatimonadetes were preferentially enriched in deeper samples, whereas Gammaproteobacteria were enriched at shallower depth (Fig. 4a, and Supplementary Table 14).

Carbon and nitrogen turnover functions in the differentially abundant genome groups also exhibited stark depth stratified patterns. C1 processing capacity and CAZy enzyme diversity were elevated in genomes more relatively abundant near the surface, while inorganic nitrogen turnover functions were enriched in genomes more relatively abundant in deeper soil (Fig. 4b, 4c, and Supplementary Tables 15-16)**.** We note that all Archaea had very low CAZy diversity, thus we conducted a separate CAZy diversity analysis with Archaea removed. This

221 additional analysis of only bacterial genomes indicates that genomes with higher relative
222 abundance at depth still harbor a significantly reduced CAZy diversity compared to genomes
223 more relatively abundant near the surface (Supplementary Fig. 11).
224
225 **Extended rainfall decreases soil depth-based functional stratification.** Sample sets
226 collected from 10-20 cm and 30-40 cm were analyzed for rainfall extension effects separately, to
227 control for the strong phylogenetic and metabolic signal observed with depth. In response to
228 rainfall extension, at 10-20 cm, 101 microorganisms increased and 72 microorganisms
229 decreased in abundance respectively (Supplementary Table 5). At 10-20 cm, the group of
230 microorganisms increasing in abundance was enriched in Bacteroidetes whereas the group that
231 decreased in abundance was enriched in Chloroflexi. At 30-40 cm, 26 microorganisms
232 increased in abundance and 59 decreased. The group of microorganisms increasing in
233 abundance at 30-40 cm was enriched in Bacteroidetes and Verrucomicrobia whereas the group
234 that decreased in abundance was enriched in Thaumarchaeota and Bathyarchaeota. Thus, in
235 response to rainfall extension, we observe an enrichment of lineages associated with complex
236 carbon degradation at both depths and a decrease of archaeal lineages in the 30-40 cm
237 samples (Fig 4a and Supplementary Table 14).
238 Metabolic profiling showed enrichment of methanol dehydrogenase in genomes of
239 microorganisms that increased in abundance at 10-20 cm with extended rainfall (Fig 4b and
240 Supplementary Table 15). At 30-40 cm, there were statistically higher numbers of inorganic
241 carbon and nitrogen processing functions including carbon monoxide dehydrogenase, nitrilase,
242 urease, and ammonia monooxygenase in genomes of microorganisms that decreased in
243 abundance with treatment (Fig. 4b and Supplementary Table 15). However, at 30-40 cm,
244 organisms increasing in abundance in response to treatment had genomes with a statistically
245 higher CAZy enzyme diversity than those that decreased in abundance (Fig. 4c and

Supplementary Table 16). However, the CAZy diversity analysis on only bacterial genomes and found no significant difference suggesting that the extended rainfall treatment does not specifically select for microorganisms with higher CAZy diversity, but instead selects against microorganisms with very low CAZy diversity (Supplementary Fig. 11). Thus, rainfall extension appears to increase C1 processing potential closer to the soil surface while causing a decrease in inorganic carbon and nitrogen processing potential at deeper depth. However, the decreased potential for processing inorganic carbon and nitrogen at depth is accompanied by a shift towards microbes with broader complex carbohydrate degradation potential.

Discussion

We recovered genomes for >50% of detected microorganisms in a grassland soil, based on coverage (which is a measure of cells sampled) (Fig. 1a), and significantly expand the availability of genomes for soil microorganisms from poorly sampled phyla. We provide evidence that metabolic systems for processing C1 compounds were relatively abundant and phylogenetically widespread, suggesting their importance in these soils (Fig. 3a). Additionally, we identified unexpected phyla encoding inorganic nitrogen turnover functions, and show that carbon and nitrogen metabolism are highly stratified across soil depths. It is also evident that climate alteration not only shifts community composition but alters the abundance of functions for important carbon and nitrogen biogeochemical cycling reactions.

Lanthanide cofactor bearing XoxF-type methanol dehydrogenases were highly prevalent, and the only methanol dehydrogenase class identified at our site. Thus, we conclude that lanthanides can be important mediators of carbon turnover in some soils. Lanthanides are often sequestered into phosphate minerals[30,31] with low biological availability[32,33], and their acquisition likely requires strong complexation of lanthanide ions by secondary metabolites such as siderophores[34]. In a recent report analyzing a subset of genomes from this site, it was found

271 that Gemmatimonadetes, Rokubacteria, and Acidobacteria harboring large numbers of XoxF
272 sequences also exhibit extensive capacity for secondary metabolite biosynthesis[35]. Thus, we
273 suspect a link between the prevalence of lanthanide-requiring enzymes and capacity to
274 biosynthesize diverse secondary metabolites that promote mineral dissolution.
275 Finding credible type-I coxL CO dehydrogenases across many phyla supports CO as an
276 important C1 energy source in soils[36], and expands the microorganism range likely performing
277 CO oxidation. However, many coxL-like sequences identified were phylogenetically unrelated to
278 genuine type-I coxL sequences, and likely have other substrates (Supplementary Fig. 8). Many
279 molybdoprotein dehydrogenases act on small molecules like nicotinate and succinate[37,38], which
280 constitute large fractions of plant exudates[39]. Thus, we suggest these enzymes may play roles
281 in plant exudate processing and turnover in the studied soils. Future research may establish that
282 these enzymes are currently under recognized mediators of small molecule turnover in soils.
283 Our data show that Gp2 Acidobacteria, which are abundant in some soils[40], encode
284 large repertoires of CAZy enzymes (Supplementary Fig. 10), and thus may represent an
285 important and overlooked complex carbohydrate turnover sink. Additionally, the high prevalence
286 of carbohydrate esterases we detected, as well as the genomic co-occurrence of acetate
287 metabolism, suggests C1 compounds and small organic molecules are important and readily
288 available carbon currencies for diverse microorganisms. As methyl and acetyl groups are
289 common additions to many polymers[27,28], the widespread prevalence of carbohydrate esterases
290 may represent a strategy where readily available C1 and C2 carbon is accessed with minimal
291 energetic investment. This observation may explain, in part, why low molecular weight carbon
292 molecules are important currencies in this ecosystem.
293 The observation that most microorganisms encoding inorganic nitrogen turnover
294 functions only harbor single steps of these pathways (Fig. 3b) parallels a similar finding for
295 complex subsurface microbial communities[21]. Thus, both soils and sediments may be structured

296 by metabolic handoffs, leading to high degrees of inter-organism cooperativity. Additionally, the
297 identification of Gemmatimonadetes with the capacity for nitrite to nitric oxide reduction, and
298 only two genomes with N_2O processing capacity, shows denitrification in these soils differs from
299 observations in other soil types[41,42]. These differences may directly impact the release of the
300 climate-change relevant gasses N_2O and NO from this system.

301 We found that grassland soils can be highly stratified both phylogenetically and
302 functionally. Additionally, deeper soils were significantly enriched in microorganism groups that
303 are underrepresented in genomic databases. These findings have broad implications for
304 understanding soil organic matter (SOM) turnover, as it is known that deeper strata account for
305 a much larger fraction of SOM, with a much longer turnover time, than SOM in shallow soil[43].
306 Thus, the genomes reported here contribute significantly to understanding of bacteria and
307 archaea that could exert critical controls on turnover rates of carbon stored in deeper soils.

308 The enrichment of enzymes involved in complex carbon metabolism, C1, and small
309 molecule turnover in microorganisms closer to the surface (Fig. 4b-c) suggests that metabolic
310 strategies at shallow depths are structured around plant-derived exudates and complex carbon.
311 These data support the observation that SOM has significantly shorter residence time closer to
312 the soil surface[43]. In contrast, most inorganic nitrogen transformation functions are more
313 prevalent or exclusively found in microorganisms enriched at greater depth (Fig. 4b). Thus, N_2O
314 discharged to the atmosphere from Mediterranean grasslands may originate from deeper soil
315 strata.

316 Under the treatment involving extended Spring rainfall, the relative decrease of
317 microorganisms at deeper depths performing ammonia liberation and oxidation suggests a
318 mechanism by which climate change could limit nitrogen cycling and N_2O release.
319 Simultaneously, increased complex carbohydrate degradation capacity at depth could counter
320 this climate change impact by increasing CO_2 release from previously recalcitrant SOM.

However, the kinetics of CO_2 and N_2O release in response to rainfall changes, and the generality of these findings to other soils, remain uncertain. What is certain is that climate change can have a direct impact on the relative abundance and metabolic capacities of microorganisms in soil ecosystems, with potentially important impacts for trace gas release.

Methods

Study Site and Rainfall Amendment

Soil samples were collected from 3 paired 70 m^2 circular plots at the south meadow field site on the Angelo Coast Range Reserve in northern California (with permission given from APP# 27790; 39°44'21.4"N 123°37'51.0"W). One plot of each pair was part of an ongoing spring rainfall extension experiment initiated in the year 2000[19], in its 14th year at the time of sampling. The rainfall extension experiment was established based on California rainfall patterns for the upcoming 50-100 years predicted by the Hadley Center for Climate Prediction and Research and the Canadian Center for Climate Modeling and Analysis in the year 2000. For plots that were treated with extended spring rainfall, every third day for 3 months during April-June, 14-16 mm of water was added over the ambient climate, reflecting a 20% increase in mean precipitation[19]. Amended water was collected from a mountain spring above the meadow, and selected as its nitrogen and trace mineral concentrations fall within the range of concentrations for natural rainwater at the site.

Edaphic factors from the soil plots sampled have been previously reported with respect to depth and spring rainfall extension treatment in detail [44]. Briefly, the sampled soils are composed of roughly 50% sand, 30% silt, and 20% clay and pH ranged between 5.34 - 5.68. Carbon concentration and C:N ratio significantly decrease with depth and significantly increased under extended spring rainfall conditions. Carbon concentration was 18 mg g-1 (10 cm) - 6 mg g-1 (50 cm) under control conditions and 26 mg g-1 (10 cm) - 14 mg g-1 (50 cm) under

346 extended rainfall. The C:N ratio was 11.2 (10 cm) - 8.1 (50 cm) under control conditions and
347 12.4 (10 cm) - 10.8 (50 cm) under extended rainfall. These measurements are also in line with
348 recently conducted pH and soluble organic carbon measurements taken from untreated soil
349 between 10-40 cm depth at the north end of the meadow[14].
350
351 **Sampling and DNA Extraction**
352 Samples were collected on five separate days beginning in September 2014 and ending
353 in October 2014, before and following fall rainfall events, as detailed in Supplementary Fig. 1
354 and Supplementary Table 1. Collection was undertaken across three biological replicate paired
355 plots at the south end of the meadow. Each plot pair consisted of one biological control plot and
356 one plot that was amended with extended spring rainfall, as detailed in Suttle et al. [19]. Prior to
357 starting a sample borehole, leaf litter and surface plant biomass were cleared from the sampling
358 location. Subsequently, we used a manual soil coring device containing a sterilized 1.5 in x 7 in
359 cylindrical polycarbonate insert to remove 10 cm of soil at a time from an individual sampling
360 bore. Soil from the first 0-10 cm of each bore was discarded, and each subsequent 10 cm soil
361 fraction was collected with a fresh sterilized insert. In the field after collection, the soil was
362 immediately removed from the insert, homogenized, put into sterile bags, and flash frozen on a
363 mixture of dry ice and ethanol. In total 60 soil samples were collected across all depths, plots,
364 and treatments. Soil samples were then maintained at -80°C prior to DNA extraction.
365 For each sample, DNA was extracted from 10 g of homogenized soil using the
366 PowerMax Soil DNA isolation kit (MoBio Laboratories Inc., Carlsbad, CA, USA) as described
367 previously[14]. Metagenomic library preparation and DNA sequencing were performed at the Joint
368 Genome Institute. Metagenomic libraries were prepared for sequencing on an Illumina
369 HiSeq2500 platform, producing 250 bp paired-end reads with a target inter-read spacing of 500

bp. Raw sequencing data were subsequently processed with the Illumina CASAVA pipeline version 1.8.

Metagenomic Assembly

Raw reads were initially assessed for quality using the FastQC analysis suite (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). FastQC analysis indicated that for some samples a > 1% GC bias existed in the last 50 bp of reads, thus reads were all initially hard trimmed to a maximum of 200 bp using BBduk (https://sourceforge.net/projects/bbmap/) with the following parameters: forcetrimright=200. Hard trimmed reads were then processed to remove Illumina adaptor sequences and phiX sequence contamination using BBduk with default parameters. Finally, reads were quality trimmed with Sickle using default parameters (https://github.com/najoshi/sickle).

The 60 samples were individually de novo assembled on a 24-core Intel Xenon Linux cluster node with 256 Gb of RAM using IDBA-UD v1.1.1[45] with the following initial parameters: --pre_correction --mink 30 --maxk 200 --step 10. In the 12 cases where assemblies did not complete due to memory requirements, minimum k-mer size was increased to 40 bp and step size was increased to 20 bp. In the 14_0903_13_30cm sample where these parameters still did not allow the assembly to complete due to memory requirements, assembly was performed using megahit with the following parameters: --k-min 41 --k-max 201 --k-step 20 --min-contig-len 1000. The contigs resulting from the megahit assembly were then scaffolded using the IDBA-UD scaffolder with the following parameters: --seed_kmer 100 --min_contig 1000. Sequencing coverage of each contig was calculated by mapping raw reads back to assemblies using Bowtie2[46]; also, see Supplementary Table 1.

Metagenome Annotation

Following metagenome assembly, all samples were filtered to remove contigs smaller than 1 kb using pullseq (https://github.com/bcthomas/pullseq). Open reading frames (ORFs) were then predicted on all contigs using Prodigal v2.6.3[47] with the following parameters: -m -p meta. Predicted ORFs were initially annotated using USEARCH[48] to search all predicted ORFs against Uniprot[49], Uniref90, and KEGG[25]. 16S ribosomal rRNA genes were predicted using the 16SfromHMM.py script from the ctbBio python package using default parameters (https://github.com/christophertbrown/bioscripts). Transfer RNAs were predicted using tRNAscan-SE[50]. The full metagenome samples and their annotations were then uploaded into our in-house analysis platform, ggKbase, where they are publically available (https://ggkbase.berkeley.edu). Please note that it is necessary to register as a user (provide an email address) to access the data.

rpS3 Identification, Clustering, and Diversity Analysis

rpS3 marker sequences were identified across all metagenomes using a custom Hidden Markov Model (HMM) based on an alignment of rpS3 sequences from the Hug, et al. tree of life dataset[51]. Briefly, all rpS3 sequences provided in Hug, et al. were initially filtered to remove Eukaryotic sequences. Sequences were then clustered at 90% ID using USEARCH with the following parameters: usearch -cluster_fast rpS3_sequences.faa -sort length -id 0.90 -maxrejects 0 -maxaccepts 0 -centroids rpS3_sequences_NR90.faa. The non-redundant sequences were then filtered to remove sequences < 200 aa in length with pullseq. The resulting set of 2,249 sequences were aligned using muscle[52] and an HMM was constructed from the alignment using HMMER3 with default parameters[53]. The HMM was benchmarked against the Uniprot reference proteomes database, and it was determined that rpS3 sequences could be confidently identified above a cutoff HMM alignment score of 40.

Across all metagenomes we identified a total of 10,159 rpS3 sequences that passed our HMM score threshold of 40. We clustered these sequences at 99% ID using USEARCH to obtain groups that roughly equate to species. We refer to these as Species Groups (SGs). The following USEARCH options were used: -cluster_fast all_rpS3.fa -sort length -id 0.99 -maxrejects 0 -maxaccepts 0 -centroids all_rpS3_centroids.faa. Subsequently we identified the longest contig in each rpS3 protein cluster to serve as a mapping target for abundance quantification of each SG (Supplementary Table 3 and Supplementary Data 2).

The longest contig representing each SG was mapped against the reads of each sample using Bowtie2 with default parameters. Mapped reads were filtered to remove all paired reads that mapped with < 99 % ID in either read pair. Reads mapped per contig were then counted to produce a read count table (Supplementary Table 3), and per base pair coverage was calculated to produce a coverage table (Supplementary Table 4). The coverage table was then normalized to the sequencing depth of each sample with the following formula: ((coverage / reads sequenced in sample) x 100,000,000) (Supplementary Table 5). For purposes of quantifying the number of detected SGs per sample we considered an SG to be present if ≥ 2 reads were mapped to its longest contig at the 99 % ID threshold.

To produce a collectors curve, we randomly selected from 1 to 60 samples without replacement using 100 sampling iterations at each sampling size. The number of unique SGs actually assembled (not just detected) in the sample subsets was quantified. We then fit a self-starting lomolino model[54] to the data using the vegan package in R[55]. From this model fit we determined the slope of the collectors curve at 60 samples as well as extrapolated the total number of SGs and number of additional SGs per sample we would recover had we doubled our sampling efforts to 120 samples over the same sample set (Supplementary Fig. 3D and 3E). Using the unfiltered read count table as input we also calculated species richness estimators

(Supplementary Fig. 3C), including the iChao2 metric[56], with the SpadeR package in R (https://github.com/AnneChao/SpadeR/).

rpS3 SGs were classified at the Phylum and Class level (where possible) by constructing a phylogenetic tree containing our sequences and rpS3 reference sequences from Hug et al. Briefly, our 3,325 representative rpS3 sequences were concatenated with a set of 2,324 reference rpS3 sequences from Hug et al. and aligned using muscle[52]. The resulting alignment was stripped of columns containing > 95% gap positions and a phylogenetic tree was constructed from the alignment using FastTree[57]. Sequences were then manually assigned Phylum and Class level lineage information based on their position relative to reference sequences in the tree.

Ordination and Variable Importance Analysis

All Ordination and variable importance analysis was performed in R using the vegan and phyloseq packages[55,58]. SG coverage values were Hellinger standardized, and then SGs were removed that had a coefficient of variation (CV) of normalized coverage > 3 or with < 5 samples where raw coverage was ≥ 0.25. A maximum likelihood phylogenetic tree for weighted Unifrac (wUnifrac) was produced from a muscle alignment of all rpS3 SG centroids using IQ-TREE[59] (Supplementary Data 4). The phylogenetic tree and normalized coverage table were then loaded into phyloseq where wUnifrac distance was calculated using the UniFrac command in phyloseq with the following parameters: weighted = TRUE, normalized = TRUE. Non-metric multidimensional scaling (NMDS) ordinations were constructed from wUnifrac distance using the metaMDS command in vegan with the following options: k = 2, try = 500, trymax = 500 (NMDS stress = 0.055). Ordinations were plotted in R using ggplot[60]. The importance of metadata variables on community composition was calculated from wUnifrac distances using the mrpp command in vegan with the following options: permutations = 10000, weight.type = 1.

Differential Abundance Analysis

Differential abundance of SGs across sampling depth and between treatment and control conditions was determined using raw read count data as the input (Supplementary Table 3) for the DEseq2 package in R[61]. We did not filter count data as DEseq filters low count data, and explicitly requests unfiltered data to more accurately estimate sample size factors and negative binomial model dispersion. To avoid linear combinations between DEseq model terms, paired plots from the same biological replicate were combined into a variable called "replicate" (plot 2 & plot 5 = A; plot 9 & plot 12 = B; plot 13 & plot 16 = C).

Differential abundance of SGs across depth was tested by comparing a full DEseq model (design = ~ Plot + Time_Point + Treat_Control + Depth) against a reduced DEseq model, where depth was omitted as a variable (reduced = ~ Replicate + Time_Point + Treat_Control), using the likelihood ratio test (LRT). Result p-values from the LRT were then corrected using the Benjamini & Hochberg procedure via false discovery rate estimation (FDR) [62], and filtered to remove results with FDR > 0.05. We then fit individual linear models to the log normalized counts of each SG showing a significant relationship with depth to determine an overall increase or decrease across the depth series. Models were fit using the lm function in R with the following form: log_count ~ depth. Model slopes and slope p-values were subsequently extracted. Slope p-values were corrected using FDR and values with FDR > 0.05 were removed. SGs with significant positive and negative model slopes were considered to increase and decrease with depth respectively.

Differential abundance between treatment and control plots was analyzed individually for each depth due to the extremely strong stratifying effect of depth. To contrast treatment and control at individual depths, a combined treatment-depth variable was created called "Factor" (ie: treatment20cm vs control20cm). A DEseq object was constructed using the following form:

493 ~ Replicate + Time_Point + Factor. DEseq was then run using the standard negative binomial
494 Wald test for GLM fits with the following options: fitType = “local”. Results where treatment and
495 control conditions were contrasted for each depth were extracted and filtered to remove SGs
496 with FDR > 0.05.
497
498 **Proteomics Methods, Annotation, and Analysis**
499 A representative subset of 20 of our soil samples were selected for full proteome
500 analysis, which was performed at the Oak Ridge National Laboratory (Oak Ridge, TN, USA). To
501 be as representative as possible samples were selected from the deepest and shallowest
502 depths sampled, from the two most geographically separated soil plots, from control and
503 extended rainfall treated plots, and from sampling dates that occurred before and after rainfall
504 events (Supplementary Figure 1).
505 Proteins were extracted from each soil sample by using the same method as described
506 previously[14]. Briefly, for each soil sample, the NoviPure Soil Protein Extraction Kit (MoBio
507 Laboratories Inc.) was used to extract proteins from 10 g of soil. A crude protein extract was
508 concentrated from 12 ml to 1 ml by using a 30 KDa Amicon® Ultra-4 Centrifugal Filter Unit
509 (Millipore). Proteins were then precipitated by trichloroacetic acid (Sigma-Aldrich) overnight at 4
510 °C and pelleted by centrifugation. Protein pellets were washed with ice-cold acetone (Sigma-
511 Aldrich) three times and re-suspended in 6 M guanidine (Sigma-Aldrich). Protein concentrations
512 were estimated using a bicinchoninic acid assay (Thermo Scientific). Fifty microgram of proteins
513 were further processed and digested using the Filter-aided Sample Preparation[63,64]. Peptides
514 were measured by a 11-step Multidimensional Protein Identification Technology[65], as described
515 previously[14,63]. MS/MS spectra from each soil sample were searched using Sipros Ensemble[66]
516 against the matched protein database constructed from the metagenome of that sample. Raw
517 search results were filtered to achieve 1% FDR at the peptide level, estimated by the target-

518 decoy approach[67]. Proteins were inferred from identified peptides using a parsimony rule[68]. A
519 minimum of one unique peptide was required for each identified protein or protein group. FDR at
520 the protein level of each sample was below 3%.

521 Proteins confidently identified in each sample were annotated using hmmsearch against
522 the dbCAN v6 HMM database with default parameters[69]. The results were filtered to remove hits
523 with an e-value ≥ 1×10^{-14} and HMM coverage ≤ 0.35. For CAZy domains overlapping the same
524 region of sequence, the domain with the lower e-value was selected. Carbon and nitrogen
525 metabolic functions were annotated by using HMMER3 against an in house HMM database built
526 from the Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology groups (See Genome
527 Metabolic Annotation below for further information). For methanol dehydrogenase (xoxF) and
528 CO dehydrogenase (coxL) genes we determined subgroup membership using initial HMM
529 placement and subsequent phylogenetic classification as described below in our methods on
530 Genome Metabolic Annotation. All annotations were concatenated and a protein that received
531 confident annotations from more than one database was assigned the annotation with the
532 highest e-value score.

533 To enable comparison across samples, proteins from each sample were clustered into
534 their assigned functional orthology groups, and the spectral counts for proteins in the same
535 sample with the same functional assignment were summed. Prior to performing further
536 statistical analysis, functions that were present in < 5 samples were removed from the analysis.
537 Subsequently, the remaining functions were ranked in each sample based on the total spectral
538 counts assigned to a function, and the mean rank for a function was calculated across samples.

539 To look at overrepresentation of KEGG functions in our proteomic dataset, we compared
540 the total number of proteins in our dataset annotated with a KEGG KO to the number of proteins
541 with that KEGG KO in the KEGG database. Over-enrichment was determined using the

hypergeometric test, implemented as the phyper function in R. All hypergeometric p-values were then corrected for multiple testing using FDR.

Genome Binning, Curation, and Dereplication

Metagenome assemblies were binned into draft genomes using a dereplication and aggregation strategy using the output of multiple metagenomic binning programs. Reads from all 60 samples were mapped to contigs > 2 Kbp using Bowtie2, and a differential coverage profile for each contig across all samples was used as input for the following differential coverage binners: ABAWCA, ABAWACA2, MaxBin2, CONCOCT and MetaBAT[70-72]. The algorithm DasTool[73] was then used to select the highest quality bins from each metagenome assembly. Bins were then manually inspected through the ggKbase web server and contigs with phylogenetic signatures that significantly deviated from bin taxonomy were removed. Bin completeness and contamination were then assessed using CheckM[74] and bins were filtered based on an established metric of ≥ 70% completeness[21]. Bins were then dereplicated across samples by matching the rpS3 containing contigs in each SG to their respective bins. The bin with the highest completeness and lowest contamination associated with each SG was then selected to be a representative of that SG. This resulted in 896 bins associated with SGs (Supplementary Table 5). Scaffolding errors in the dereplicated bin set were corrected as previously described[21]. Gene loci in these bins were then re-called using Prodigal in single genome mode. Error corrected bins were assessed again for completeness and contamination with CheckM, and 793 bins passed the criteria of ≥ 70% completeness and < 10 % contamination that we required for inclusion in our metabolic analysis.

Genome Phylogenetic Classification

566 Taxonomy of microorganisms represented by the 896 dereplicated bins was determined
567 using the combination of a concatenated ribosomal protein (RP) tree, rpS3 protein tree, and 16S
568 rRNA gene sequences binned with genomes. For the RP tree we searched each genome for 15
569 RPs (L2, L3, L4, L5, L6, L14, L15, L16, L18, L24, S3, S8, S17, S19) using USEARCH against a
570 database of RPs from Hug et al. [51]. If RPs in a genome were not found in a contiguous block,
571 we manually checked if any of the RP containing contigs represented a contaminating
572 sequence. 852 genomes containing 8 or more RPs were then included in the analysis. RP
573 sequences were individually combined with reference RP sequences from Hug et al. and
574 selected sequences from Parks et al. [51,75]. Sequences were then individually aligned using
575 MAFFT. The resulting alignments were stripped of columns containing > 95% gap positions.
576 Individual stripped alignments were concatenated and a phylogenetic tree was constructed
577 using RAxML v8.2.10[76] on the CIPRES Science Gateway{Miller:vv}. RAxML was called as
578 follows: raxmlHPC-HYBRID -s input -N autoMRE -n result -f a -p 12345 -x 12345 -m
579 PROTCATLG. Genomes were then manually assigned Phylum and Class level lineage
580 information based on their position relative to reference sequences in the tree. Also, see
581 Supplementary Fig. 4 and Supplementary Data 5-7. In the case a genome was not included in
582 the RP tree, its taxonomy assigned by the rpS3 tree was inherited. For acidobacterial genomes,
583 Class level assignments were made by a combination of RP tree assignments and predicted
584 16S rRNA gene sequence taxonomy. Also see Supplementary Fig. 5.
585 16S rRNA gene sequences identified within metagenome bins (see above) were aligned
586 using SINA v1.2.11 implemented on the SILVA ACT web portal[77]. Sequences were aligned
587 against the global SILVA alignment for SSU rRNA genes, and sequences with an alignment
588 identity ≥ 70% were then classified using the least common ancestor method based on
589 taxonomies in SILVA. Also see Supplementary Table 6.
590

Genome Metabolic Annotation

For the 793 bins passing completeness and contamination criteria, carbohydrate active enzymes (CAZy) were annotated using hmmsearch against the dbCAN v6 HMM database with default parameters[69]. The results were filtered to remove hits with an e-value ≥ 1 x 10^{-14} and HMM coverage ≤ 0.35. For CAZy domains overlapping the same region of sequence, the domain with the lower e-value was selected. Carbon and nitrogen metabolic functions were annotated by using HMMER3 against an in house HMM database built from the Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology groups (KOs). Briefly, all KEGG database proteins with KOs were compared with all-v-all global similarity search using USEARCH. MCL was then used to sub-cluster KOs (inflation_value = 1.1). Each sub-cluster was aligned using MAFFT, and HMMs were constructed from sub-cluster alignments. HMMs were then scored against all KEGG sequences with KOs and a score threshold was set for each HMM at the score of the highest scoring hit outside of that HMMs sub-cluster. Access to the proprietary KEGG database was secured via contract, so only our procedure to profile them can be made public.

For methanol dehydrogenase (xoxF), CO dehydrogenase (coxL), and nitrite reductase (nirK) we constructed individual phylogenetic trees to discriminate homologous, but functionally distinct, proteins that can be identified by HMM search alone. XoxF sequences were initially identified in genomes using a custom HMM for PQQ-binding alcohol dehydrogenases[21]. Angelo sequences were combined with reference sequences from Keltjens et al. and Taubert et al. [33,78], and aligned using MAFFT. A Phylogenetic tree was constructed using FastTree (Supplementary Fig. 7 and Supplementary Data 12) and xoxF sequences were manually discriminated from mxaF and general ADH sequences by their position relative to reference sequences in the tree.

Putative coxL sequences were identified by KEGG HMM hits to K03520. Angelo hits were combined with reference sequences from Quiza et al. [9], and aligned using MAFFT. A

Phylogenetic tree was constructed using FastTree (Supplementary Fig. 8 and Supplementary Data 13) and coxL-typeI sequences were manually identified by a known sequence motif "AYRCSFR" [22] and their position relative to reference sequences in the tree.

Putative nirK sequences were identified by KEGG HMM hits to K00368. Angelo hits were combined with reference sequences from Decleyre et al. [79], and aligned using MAFFT. A phylogenetic tree was constructed using FastTree (Supplementary Fig. 9 and Supplementary Data 14), and true NirK sequences were manually identified by the presence of properly aligned catalytic residues and their position relative to reference sequences in the tree.

C1 Carbon and inorganic Nitrogen metabolism were assessed by by looking at a specific set of 28 targeted functions. For further information on annotation criteria and functional assignments to genomes see Supplementary Tables 9-13.

Depth and Treatment Enrichment Analysis

We first assessed the differences between estimated completeness and contamination for the sets of genomes that would be compared when testing for enrichment (Supplementary Fig. 14A, 14B, and Supplementary Table 19). For each condition tested (Depth, Treatment - 20cm, and Treatment - 40 cm), the estimated genome completeness and contamination values across the three response groups (Increase, Decrease, and Neither) was initially tested for significant differences using the Kruskal-Wallis rank sum test[80], implemented as the kruskal.test function in R. The Kruskal-Wallis p-values were corrected for multiple testing using FDR, and post-hoc testing between specific response groups was undertaken for FDR ≤ 0.1. Post-hoc testing was carried between all pairs of response groups in a condition using the Wilcoxon Rank Sum test implemented as the pairwise.wilcox.test function in R[81,82], and corrected for multiple testing using FDR. An FDR ≤ 0.05 in post-hoc testing was considered significant.

Significant enrichments of phylum level lineages and 29 targeted metabolic functions were assessed between genome response groups in each condition using Fisher's exact test[83] followed by post-hoc testing with a permutation analysis (Supplementary Tables 9-10 and 14-15). We first removed all metabolic functions or phyla from the analysis that were not present in both the increased or decreased genome groups from a condition. Then, for each factor tested across the three conditions (Depth, Treatment - 20cm, and Treatment - 40 cm) counts in the three genome response groups (Increase, Decrease, and Neither) were first compared using Fisher's exact test[83] on a 2X3 contingency table, implemented as the fisher.test function in R. Fisher test p-values were corrected using FDR, and post-hoc testing was carried out on functions or phylum categories with FDR ≤ 0.1. Post-hoc testing was then only conducted on groups of genomes that increased or decreased with respect to a condition. We carried out post-hoc testing using a permutation test implemented as a custom R function to reflect the underlying frequency distribution of the phylum or functional gene being tested across all 793 bins that were analyzed (See Code Availability Statement). Briefly, the counts of each function or phylum in the increased or decreased sets of genomes were randomly resampled without replacement 10,000 times from all 793 genomes. The absolute value of the difference between the fraction of counts of a phylum or function over the total number of genomes in the respective increased or decreased set was then calculated. P-values were calculated as the number of absolute fractional differences in the permuted set that exceeded the observed fractional difference divided by 10,000 samples. P-values were corrected using FDR, and FDR values ≤ 0.05 were considered significant.

CAZy enzyme Shannon and Simpson diversity for genomes was quantified using the diversity function in the R vegan package[55]. Unique counts of CAZy enzymes per genome were quantified with the specnumber function in the R vegan package. For each diversity metric calculated across the three conditions (Depth, Treatment - 20cm, and Treatment - 40 cm), the

three genome response groups (Increase, Decrease, and Neither) were first compared using the Kruskal-Wallis rank sum test[80], implemented as the kruskal.test function in R. Kruskal-Wallis p-values were corrected using FDR, and post-hoc testing between groups of genomes that increased and decreased with respect to a condition, were conducted for all FDR ≤ 0.1. Post-hoc testing was carried out using the Wilcoxon rank-sum test implemented as the wilcox.test function in R[81,82], and corrected for multiple testing using FDR. Differences were considered significant for FDR values ≤ 0.05. Differential enrichment of specific CAZy classes was tested using the same procedure as for diversity metrics, with initial three category testing being performed with the Kruskal-Wallis test, subsequent post hoc testing between increased and decreased genomes being performed with the Wilcoxon rank-sum test, and multiple testing being corrected with FDR.

Feature selection of KEGG KOs that were significant predictors of a genome having increased or decreased abundance with depth was undertaken using the random forest-based method Boruta, implemented in R[84]. Briefly, KO profiles from genomes showing a depth response were subset and KOs present in ≤ 5 genomes of the total set were removed from the data. Due to the significantly different number of genomes that increase and decrease with depth case weights were applied based on the ratio of increasing to decreasing genomes (Decrease = 2.184, Increase = 1). Brouta was then called with the following options: Depth_Change ~ ., doTrace = 2, maxRuns = 500, num.trees = 7500, case.weights = cs_wts. All features confirmed as significant predictors by Boruta were then individually tested for differential abundance between the genome sets that decreased and increased with depth using the Wilcoxon rank-sum test. Wilcoxon p-values were FDR corrected, and KOs with an FDR ≤ 0.05 were considered significant. For full Boruta output see Supplementary Table 18.

Functional Gene Co-Occurrence and Correlation Analysis

The co-occurrence overview of 29 targeted carbon and nitrogen turnover functions annotated in our 793 genomes (Supplemental Fig. 12 and Supplementary Table 10) was produced using the pheatmap function in R[82]. Clustering was performed using binary distance and Ward hierarchical grouping[85]. Correlations and correlation p-values for the co-occurrence of functions across all genomes (Supplemental Fig. 13) were calculated using spearman rank correlation implemented with the rcorr function in R[82]. All p-values were corrected using FDR, and FDR values ≤ 0.05 were considered significant. Significant correlations between functional genes were plotted using the corrplot function from the corrplot package in R (https://github.com/taiyun/corrplot). Correlations were clustered using the angular order of the eigenvectors implemented in the corrplot package, and cluster groups were human defined.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data Availability

Genomic data including curated genomes and raw sequencing reads are available under the NCBI BioProject accession number PRJNA449266. Proteomic data are available through the ProteomeXchange Consortium via the PRIDE partner repository with identifier PXD013110.

Code Availability

Code used in the analysis for this paper are available at the following GitHub repository: https://github.com/SDmetagenomics/Angelo2019_Paper

Main Figure Legends

Figure 1 l rpS3 species group abundance, influence of variables, and abundance metrics. (**A**) Percent of total coverage of all Species Groups (SGs) ranked by relative phylum coverage. "Other" includes phyla with < 5 SGs. Organisms in red are in the top 25% of organisms by coverage. The inset pie chart shows the breakdown of SGs associated with genome bins (blue) based on count and coverage of SGs. (**B**) NMDS plot (stress = 0.055) of SG UniFrac distances. The ordination is replicated and overlaid with the four data types collected across our 60 samples. Variable importance (C) and significance (p) calculated by Multi Response Permutation Procedure (MRPP) are displayed in the legend. (**C**) Top 25% of SGs ranked by total coverage across all samples. The inset shows the full rank abundance curve and shows the positions where 25%, 50%, and 75% of the total dataset coverage are reached. Red ticks under the plot indicate SGs with bins. Also see Supplementary Table 5.

Figure 2 l Maximum likelihood tree of all near complete genomes. Phylogenetic tree constructed with a concatenated alignment of 15 co-located ribosomal proteins (L2, L3, L4, L5, L6, L14, L15, L16, L18, L24, S3, S8, S17, S19). Tree includes 722 Bacterial and 71 Archaeal genomes. The two Chloroflexi classes basal to classic Chloroflexi lineages are named. Concentric rings moving outward from the tree indicate if a genome's associated SG abundance was found to significantly increase or decrease with depth and increase or decrease in plots under extended rainfall treatment at either 10-20 cm or 30-40 cm. For all genomes shown, the direction of response (increase or decrease) to extended rainfall treatment was never different between depths. The concentric bar plot indicates relative abundance (Methods). For the complete ribosomal protein tree, see Supplementary Fig. 4, and Supplementary Data 5. For all exact relative abundance values and differential abundance statistics see Supplementary Table 5.

Figure 3 l Predicted carbon and nitrogen metabolic transformations. (**A**) Predicted phylum level genomic capacity for breakdown of small carbon- and nitrogen-containing compounds, and liberation of methyl and acetyl groups from complex polymers. Horizontal bar plots indicate the fraction of genomes within a phylum encoding each function (as shown in the key on the bottom left). Numbers to the right of

742 bars in parentheses indicate the total genes detected (n = 793 independent genomes). (**B**) Counts of
743 genomes encoding capacities for individual or multiple nitrogen transformation steps. AMON = Ammonia
744 Oxidation to Nitrate; NRA = Nitrate Reduction to Ammonia; DNIT = Denitrification (n = 793 independent
745 genomes). (**C**) Top panel: counts of carbohydrate active (CAZy) enzymes across genomes in each
746 phylum. Points indicate the total counts in individual genomes and point sizes reflects genome relative
747 coverage across all samples (as shown in the key on the bottom left). Box plots enclose 1st to 3rd quartile
748 of data values, with a black line at the median value. Top inset: Bar plot showing the total number of
749 CAZy enzymes across all genomes belonging to each CAZy class (GH = glycosyl hydrolase; CE =
750 carbohydrate esterase; AA = auxiliary activity; PL = polysaccharide lyase). Bottom panel: the count of all
751 246 possible CAZy enzymes types that were identified across a phylum is shown (n = 793 independent
752 genomes). Also see Supplementary Tables 10-13.
753
754 **Figure 4 | Enrichment of phyla and metabolic functions across depth and treatment.** (**A**) The
755 difference in proportion of a phylum between genome groups that increase and decrease with
756 depth/rainfall extension. Black stars indicate a significant enrichment of the phylum and bar direction
757 indicates the genome set where the enrichment was found (two-sided permutation test; * FDR ≤ 0.05, **
758 FDR ≤ 0.01, *** FDR ≤ 0.001) (**B**) The count of genomes encoding targeted carbon and nitrogen
759 processing functions found to be significantly enriched in at least one comparison between genome
760 groups that increase and decrease with depth/rainfall extension treatment. Genome counts only include
761 those that were statistically different between depth or treatment shown. Black stars indicate a significant
762 enrichment of the function and bar direction indicates the genome set where the enrichment was found
763 (two-sided permutation test; * FDR ≤ 0.05, ** FDR ≤ 0.01, *** FDR ≤ 0.001). Colors indicate phyla (see
764 Fig. 3). Nitrilase (nit); urease (URE); methylamine dehydrogenase (mauAB); ammonia monooxygenase -
765 particulate monooxygenase (amo-pmo); nitrite oxioreductase (nxrAB); nitrite reductase (nirK); methanol
766 dehydrogenase (xoxF); formaldehyde oxidation - direct (Fal1); formaldehyde oxidation - glutathione
767 pathway (Fal2); formate oxidation – multi-subunit (Frm2); carbon monoxide dehydrogenase (coxL-I). (**C**)
768 CAZy enzyme Simpson diversity distributions between genome groups that increase and decrease with

depth/rainfall extension treatment. Simpson diversity has been transformed to the inverse form (1/(1-simpson)) for ease of viewing. Points are colored by phylum (see Fig. 3). A black star between box plots indicates a statistically difference (two-sided Wilcoxon test; * FDR ≤ 0.05). Across all figure panels sample numbers were: n_{depth} = 60 biologically independent samples, $n_{20cm\_treatment}$ = 24 biologically independent samples, $n_{40cm\_treatment}$ = 20 biologically independent samples. Across all figure panels the number of genomes analyzed were: n_{depth} = 570 independent genomes, $n_{20cm\_treatment}$ = 173 independent genomes, $n_{40cm\_treatment}$ = 85 independent genomes. All tests were corrected for multiple testing using false discovery rate (FDR). For all exact FDR values see Supplementary Tables 14-16.

Acknowledgements

We thank Sue Spaulding for assistance with fieldwork and Evan Starr for helpful discussions on data analysis and figure production. Sequencing was carried out under a Community Sequencing Project at the Joint Genome Institute. Funding was provided by the Office of Science, Office of Biological and Environmental Research, of the US Department of Energy Grant DOE-SC10010566

Author Contributions

SD, PA, ZL, CP, TN and JFB conceived analysis. CP, TN and JFB designed sampling strategy. DB performed soil sampling. SD and BCT performed genomic sequence processing and assembly. SD, PA, ACC, DB, KA, KRL, and BCT performed annotation and parsing genomic data. ZL and CP performed proteomics. SD, PA and ACC performed statistical analysis on genomic and proteomic datasets. SD and JFB wrote manuscript. SD, PA, ACC, CP, TN, and JFB edited manuscript.

Competing Interests

The authors declare no competing interests.

Corresponding Author

All correspondence should be addressed to Jillian F. Banfield (jbanfield@berkeley.edu)

References

1. Boval, M. & Dixon, R. M. The importance of grasslands for animal production and other functions: a review on management and methodological progress in the tropics. *Animal* **6,** 748–762 (2012).
2. Eze, S., Palmer, S. M. & Chapman, P. J. Soil organic carbon stock in grasslands: Effects of inorganic fertilizers, liming and grazing in different climate settings. *J. Environ. Manage.* **223,** 74–84 (2018).
3. Conrad, R. Soil microorganisms as controllers of atmospheric trace gases (H-2, CO, CH4, OCS, N2O, and NO). *Microbiol. Rev.* **60,** 609–+ (1996).
4. Gougoulias, C., Clark, J. M. & Shaw, L. J. The role of soil microbes in the global carbon cycle: tracking the below-ground microbial processing of plant-derived carbon for manipulating carbon dynamics in agricultural systems. *J. Sci. Food Agric.* **94,** 2362–2371 (2014).
5. Delgado-Baquerizo, M. *et al.* A global atlas of the dominant bacteria found in soil. *Science* **359,** 320–325 (2018).
6. Fierer, N. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat Rev Micro* 1–12 (2017). doi:10.1038/nrmicro.2017.87
7. Bahram, M. *et al.* Structure and function of the global topsoil microbiome. *ISME J* 1–24 (2018). doi:10.1038/s41586-018-0386-6

8. Alves, R. J. E. *et al.* Nitrification rates in Arctic soils are associated with functionally distinct populations of ammonia-oxidizing archaea. **7,** 1620–1631 (2013).

9. Quiza, L., Lalonde, I., Guertin, C. & Constant, P. Land-use influences the distribution and activity of high affinity CO-oxidizing bacteria associated to type I-coxL genotype in soil. *Front Microbiol* **5,** 1–15 (2014).

10. Barber, N. A., Chantos-Davidson, K. M., Amel Peralta, R., Sherwood, J. P. & Swingley, W. D. Soil microbial community composition in tallgrass prairie restorations converge with remnants across a 27-year chronosequence. *Environmental Microbiology* **19,** 3118–3131 (2017).

11. Cong, J. *et al.* Analyses of soil microbial community compositions and functional genes reveal potential consequences of natural forest succession. *Sci Rep* **5,** 1–11 (2015).

12. Ben J Woodcroft *et al.* Genome-centric view of carbon processing in thawing permafrost. *ISME J* 1–24 (2018). doi:10.1038/s41586-018-0338-1

13. Ji, M. *et al.* Atmospheric trace gases support primary production in Antarctic desert surface soil. *ISME J* **552,** 400–403 (2017).

14. Butterfield, C. N. *et al.* Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone. *PeerJ* **4,** e2687–28 (2016).

15. Howe, A. C. *et al.* Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 4904–4909 (2014).

16. Delmont, T. O. *et al.* Reconstructing rare soil microbial genomes using in situ enrichments and metagenomics. *Front Microbiol* **6,** 1–15 (2015).

17. Placella, S. A. & Brodie, E. L. Rainfall-induced carbon dioxide pulses result from sequential resuscitation of phylogenetically clustered microbial groups. in (2012). doi:10.1073/pnas.1204306109/-/DCSupplemental

18. Blazewicz, S. J., Schwartz, E. & Firestone, M. K. Growth and death of bacteria and fungi underlie rainfall-induced carbon dioxide pulses from seasonally dried soil. *Ecology* **95,** 1162–1172 (2014).

19. Suttle, K. B., THOMSEN, M. A. & Power, M. E. Species interactions reverse grassland responses to changing climate. *Science* **315,** 640–642 (2007).

20. Sharon, I. *et al.* Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. *Genome Research* **25,** 534–543 (2015).

21. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun* **7,** 13219 (2016).

22. Lalonde, I. & Constant, P. Identification of Unknown Carboxydovore Bacteria Dominant in Deciduous Forest Soil via Succession of Bacterial Communities, coxLGenotypes, and Carbon Monoxide Oxidation Activity in Soil Microcosms. *Applied and Environmental Microbiology* **82,** 1324–1333 (2016).

23. Weber, C. F. & King, G. M. Quantification of Burkholderia coxL Genes in Hawaiian Volcanic Deposits. *Applied and Environmental Microbiology* **76,** 2212–2217 (2010).

24. Huang, L. *et al.* dbCAN-seq: a database of carbohydrate-active enzyme (CAZyme) sequence and annotation. *Nucleic Acids Res.* **46,** D516–D521 (2017).

25. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).

26. Kielak, A. M., Barreto, C. C., Kowalchuk, G. A., van Veen, J. A. & Kuramae, E. E. The Ecology of Acidobacteria: Moving beyond Genes and Genomes. *Front Microbiol* **7,** 1674–16 (2016).

27. Biely, P. Microbial carbohydrate esterases deacetylating plant polysaccharides. *Biotechnology Advances* **30,** 1575–1588 (2012).

28. Nakamura, A. M., Nascimento, A. S. & Polikarpov, I. Structural diversity of carbohydrate

esterases. *Biotechnology Research and Innovation* **1,** 35–51 (2017).

29. Eichorst, S. A. *et al.* Genomic insights into the Acidobacteriareveal strategies for their success in terrestrial environments. *Environmental Microbiology* **20,** 1041–1063 (2018).

30. Taunton, A. E., Welch, S. A. & Banfield, J. F. Microbial controls on phosphate and lanthanide distributions during granite weathering and soil formation. *Chemical Geology* **169,** 371–382 (2000).

31. Banfield, J. F. & EGGLETON, R. A. Apatite Replacement and Rare-Earth Mobilization, Fractionation, and Fixation During Weathering. *Clays and Clay Minerals* **37,** 113–127 (1989).

32. Hibi, Y. *et al.* Molecular structure of La3+-induced methanol dehydrogenase-like protein in Methylobacterium radiotolerans. *Journal of Bioscience and Bioengineering* **111,** 547–549 (2011).

33. Keltjens, J. T., Pol, A., Reimann, J. & Op den Camp, H. J. M. PQQ-dependent methanol dehydrogenases: rare-earth elements make a difference. *Appl Microbiol Biotechnol* **98,** 6163–6183 (2014).

34. Christenson, E. A. & Schijf, J. Stability of YREE complexes with the trihydroxamate siderophore desferrioxamine B at seawater ionic strength. *Geochimica et Cosmochimica Acta* **75,** 7047–7062 (2011).

35. Crits-Christoph, A., Diamond, S., Butterfield, C. N., Thomas, B. C. & Banfield, J. F. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. *ISME J* **558,** 440–444 (2018).

36. Weber, C. F. & King, G. M. Distribution and diversity of carbon monoxide-oxidizing bacteria and bulk bacterial communities across a succession gradient on a Hawaiian volcanic deposit. *Environmental Microbiology* **12,** 1855–1867 (2010).

37. Leimkühler, S. & Iobbi-Nivol, C. Bacterial molybdoenzymes: old enzymes for new

894 purposes. *FEMS Microbiology Reviews* **40,** 1–18 (2016).
895 38. Kim, S. W., Luykx, D., deVries, S. & Duine, J. A. A second molybdoprotein aldehyde
896 dehydrogenase from Amycolatopsis methanolica NCIB 11946. *Arch. Biochem. Biophys.*
897 **325,** 1–7 (1996).
898 39. Zhalnina, K. *et al.* Dynamic root exudate chemistry and microbial substrate preferences
899 drive patterns in rhizosphere microbial community assembly. *Nature Microbiology* **3,** 470–
900 480 (2018).
901 40. Bartram, A. K. *et al.* Exploring links between pH and bacterial community composition in
902 soils from the Craibstone Experimental Farm. *FEMS Microbiology Ecology* **87,** 403–415
903 (2013).
904 41. Cardenas, E., Orellana, L. H., Konstantinidis, K. T. & Mohn, W. W. Effects of timber
905 harvesting on the genetic potential for carbon and nitrogen cycling in five North American
906 forest ecozones. *Sci Rep* **8,** 3142 (2018).
907 42. Pajares, S. & Bohannan, B. J. M. Ecology of Nitrogen Fixing, Nitrifying, and Denitrifying
908 Microorganisms in Tropical Forest Soils. *Front Microbiol* **7,** 921–20 (2016).
909 43. Cheng, L. *et al.* Warming enhances old organic carbon decomposition through altering
910 functional microbial communities. *The ISME Journal* **11,** 1825–1835 (2017).
911 44. Berhe, A. A., Suttle, K. B., Burton, S. D. & Banfield, J. F. Contingency in the direction and
912 mechanics of soil organic matter responses to increased rainfall. *Plant Soil* **358,** 371–383
913 (2012).
914 45. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for
915 single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*
916 **28,** 1420–1428 (2012).
917 46. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*
918 **9,** 357–359 (2012).

47. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11,** 119–11 (2010).

48. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461 (2010).

49. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169 (2017).

50. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).

LETTER

https://doi.org/10.1038/s41586-018-0207-y

Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis

Alexander Crits-Christoph[1,2], Spencer Diamond[3], Cristina N. Butterfield[3], Brian C. Thomas[3] & Jillian F. Banfield[2,3,4,5]*

In soil ecosystems, microorganisms produce diverse secondary metabolites such as antibiotics, antifungals and siderophores that mediate communication, competition and interactions with other organisms and the environment[1,2]. Most known antibiotics are derived from a few culturable microbial taxa[3], and the biosynthetic potential of the vast majority of bacteria in soil has rarely been investigated[4]. Here we reconstruct hundreds of near-complete genomes from grassland soil metagenomes and identify microorganisms from previously understudied phyla that encode diverse polyketide and nonribosomal peptide biosynthetic gene clusters that are divergent from well-studied clusters. These biosynthetic loci are encoded by newly identified members of the Acidobacteria, Verrucomicobia and Gemmatimonadetes, and the candidate phylum Rokubacteria. Bacteria from these groups are highly abundant in soils[5–7], but have not previously been genomically linked to secondary metabolite production with confidence. In particular, large numbers of biosynthetic genes were characterized in newly identified members of the Acidobacteria, which is the most abundant bacterial phylum across soil biomes[5]. We identify two acidobacterial genomes from divergent lineages, each of which encodes an unusually large repertoire of biosynthetic genes with up to fifteen large polyketide and nonribosomal peptide biosynthetic loci per genome. To track gene expression of genes encoding polyketide synthases and nonribosomal peptide synthetases in the soil ecosystem that we studied, we sampled 120 time points in a microcosm manipulation experiment and, using metatranscriptomics, found that gene clusters were differentially co-expressed in response to environmental perturbations. Transcriptional co-expression networks for specific organisms associated biosynthetic genes with two-component systems, transcriptional activation, putative antimicrobial resistance and iron regulation, linking metabolite biosynthesis to processes of environmental sensing and ecological competition. We conclude that the biosynthetic potential of abundant and phylogenetically diverse soil microorganisms has previously been underestimated. These organisms may represent a source of natural products that can address needs for new antibiotics and other pharmaceutical compounds.

We reconstructed draft genomes for hundreds of microorganisms from the soil ecosystem of a northern Californian grassland using genome-resolved metagenomic methods, and targeted genomes from four dominant soil phyla for analysis of their biosynthetic potential (Extended Data Fig. 1). Specifically, we analysed newly reconstructed genomes from 149 Acidobacteria, 135 Verrucomicrobia, 43 Rokubacteria and 49 Gemmatimonadetes species (Supplementary Table 1 and Supplementary Methods). We targeted these groups because bacteria from all four phyla are highly abundant at our field sampling site[8] (Fig. 1a) and in globally sampled soils[5]. Specifically, meta-analysis of many 16S rRNA gene sequence studies showed that Acidobacteria and Verrucomicrobia are the first and second most abundant bacterial phyla in soil, respectively[5], and Gemmatimonadetes are also known to be common in soils[9]. There are few reference genomes available for soil-associated bacteria from all four phyla, and their potential for secondary metabolism remains understudied. To our knowledge, the current study represents the largest genomic sampling of soil-associated bacteria from these groups to date and the most detailed analysis of their secondary metabolism.

Within the genomes, we identified 1,159 biosynthetic gene clusters on contigs at least 10 kb in length (Fig. 1b and Supplementary Table 2) and an additional 440 biosynthetic gene clusters on smaller contigs (Supplementary Table 3) using antiSMASH 3.0[10], an in silico pipeline that was originally verified against 473 verified biosynthetic gene clusters with a 97.7% reported accuracy[11]. The gene clusters that we identified are inferred to synthesize nonribosomal peptides (NRPs), polyketides, terpenes, bacteriocins, lassopeptides, lantipeptides and metabolites of uncertain function. Most known bacterial natural products—including many of the clinical antibiotics that we use today—have been obtained from microbial isolates[3] of the Actinobacteria, Proteobacteria and *Bacillus*, which represent microorganisms that often comprise a minority in soil microbial communities[4,5]. Previous global analyses based on the few publicly available genomes for Acidobacteria, Verrucomicrobia and Gemmatimonadetes[12–14] identified only a handful of biosynthetic clusters, and to our knowledge only the Acidobacteria have previously been suggested to be linked to secondary metabolite production[7,15]. We greatly expand the number of known biosynthetic gene pathways from these soil microorganisms and at the same time confidently link them to their genomic contexts.

Most previous searches for biosynthetic systems from uncultivated microorganisms have randomly cloned environmental DNA into a host organism to screen for function (functional metagenomics)[16]. Other studies[2,17] have used degenerate PCR primers to explore the genetic diversity of novel biosynthetic clusters without the need for cloning, but primers can fail to amplify genetically divergent sequences. Because we reconstructed near-complete genomes de novo, we could identify entire novel biosynthetic gene clusters as well as describe their genomic, phylogenetic and ecological contexts within individual genomes and the environment. We computationally tested the ability of sets of previously used degenerate primers[2,17] to detect genes containing polyketide ketoacyl synthase and NRP amino acid adenylation domains in the clusters reported here, and found that only 5 out of 240 clusters would be likely to amplify properly when using degenerate primers (Supplementary Table 6).

Gene clusters containing nonribosomal peptide synthetases (NRPSs) and polyketide synthases (PKSs) were of particular interest, as the products of these enzymes include many antibiotics, antifungals, siderophores and immunosuppressants[14]. These NRPS and PKS biosynthetic pathways use modular enzymatic domains to build molecules with complex chemical structures. We identified 240 NRPS, PKS (types I, II and III, which differ in the organization of their enzymatic domains) and hybrid (NRPS-PKS) gene clusters on contigs from all four phyla of interest (Fig. 1c and Supplementary Table 4) and 86 probably incomplete clusters on smaller genome fragments. Although they

[1]Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, CA, USA. [2]The Innovative Genomics Institute, University of California, Berkeley, Berkeley, CA, USA. [3]Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, CA, USA. [4]Department of Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, CA, USA. [5]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. *e-mail: jbanfield@berkeley.edu

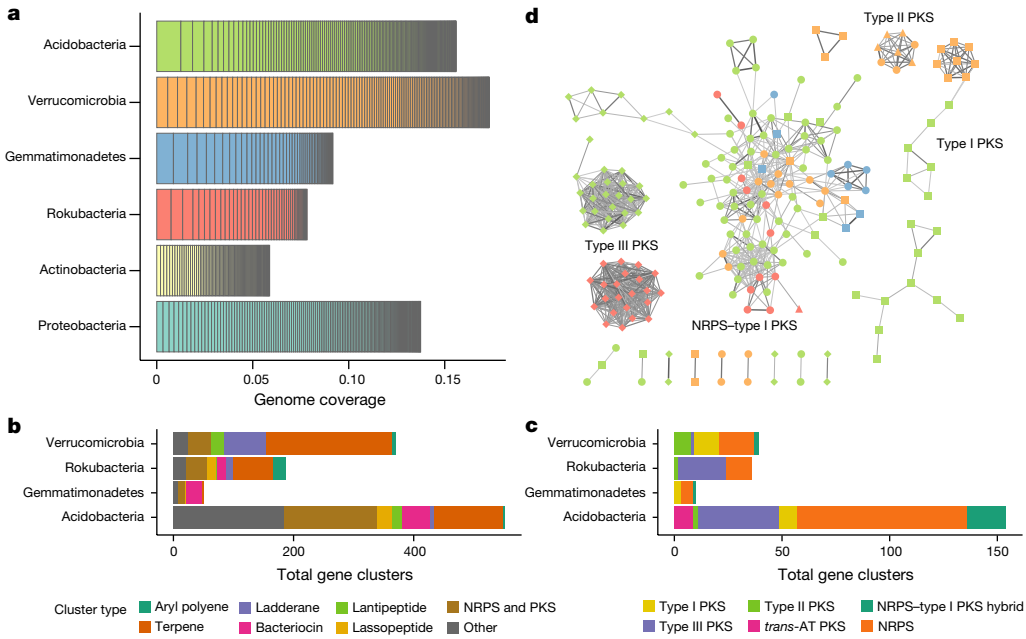


Fig. 1 | Diversity of extracted soil genomes and their biosynthetic gene clusters. a, Mean relative abundances of reconstructed genomes across 60 soil samples as determined by sequencing coverage of the genomes. Genomes from four understudied soil phyla are juxtaposed with recovered genomes from the Actinobacteria and Proteobacteria for comparison. **b**, Biosynthetic gene clusters found on contigs greater than 10 kb, from each phylum studied, coloured by putative product types as assigned by antiSMASH. **c**, NRPS and PKS gene clusters found on contigs >10 kb, from each phylum studied. **d**, Network of biosynthetic gene clusters, in which edges connect clusters that share genes. The line thickness and darkness increase with increasing percentage of genes shared between clusters. *trans*-AT, *trans*-acyltransferase.

are enormously diverse in gene content, these biosynthetic pathways are identifiable owing to their colocalized logical organization of conserved enzymatic domains. Although the majority of these clusters occurred in a wide diversity of Acidobacteria, we also identified 11 NRPS clusters in genomes of the Rokubacteria, a recently described phylum that was not previously known to produce natural products. The co-linear 'assembly-line' regulation of many NRPS and type I PKS systems make predictions of the core scaffold of the molecular product synthesized possible[11,18]. In 136 cases, there were a sufficient number of functional domains with known substrate specificity to predict the core chemical structures of the products using antiSMASH (Supplementary Table 4).

To compare the degrees to which predicted biosynthetic clusters shared genes, we built a relational network of clusters on the basis of shared gene content. This approach revealed substantial genetic variety, with large groups of diverse and sparsely connected NRPS and PKS systems in Verrucomicrobia, Acidobacteria and Rokubacteria and many unique NRPS-based clusters with few close representatives (Fig. 1d). A conserved type III PKS locus that was nearly ubiquitous in the Rokubacteria formed a dense network cluster, as did a conserved type III PKS locus found in a wide clade of the Acidobacteria. The high conservation of these type III PKS loci across taxonomic groups could indicate a broad distribution of a novel group of specialized metabolites.

We compared the 240 NRPS and PKS gene clusters to the reference set described in the 'Minimum Information about a Biosynthetic Gene' (MIBiG) repository[19] (Supplementary Table 5). No protein in any cluster shared with reference proteins more than 79.7% amino acid identity across ≥50% of the full protein lengths. Fifty-nine per cent of predicted proteins had no ≥50%-length homologue in MIBiG, and those that did shared an average of only about 39% amino acid identity to the best hit of any MIBiG protein. Using the same thresholds for gene homologues, we found that 220 clusters did not share more than 50% of the genes of any previously described cluster. Although the relationship between gene similarity of biosynthetic genes and structural similarities of their final products can be difficult to discern, previous analyses have shown that structural divergence correlates strongly with genetic divergence, even within families of gene clusters[20].

It is often the case that antibiotic producers will also encode antibiotic resistance genes to avoid self-toxicity, and that these genes will often co-localize with the antibiotic biosynthetic cluster in the genome[21]. Therefore, the presence of antimicrobial resistance genes within a gene cluster could indicate that the cluster is involved in antibiotic production. We mined all NRPS and PKS biosynthetic loci with a set[22] of curated hidden Markov models for antibiotic resistance proteins (in part derived from the Resfams[23] database) (Supplementary Methods). One hundred and fifty-three proteins from 84 different NRPS and PKS clusters most closely matched hidden Markov models for transporters known to be involved in antimicrobial resistance, out of a total of 621 transporter genes within clusters. Annotations that could most confidently be linked to antibiotic resistance included one D-alanine–D-alanine ligase in a Rokubacteria NRPS cluster, four D-alanine–D-alanine ligases in acidobacterial NRPS clusters, and two modified penicillin-binding protein sequences in Verrucomicrobia NRPS clusters (Supplementary Table 7).

Two near-complete genomes of divergent Acidobacteria were found to encode unusually large repertoires of NRP and PKS gene clusters. We refer to these two organisms as '*Candidatus* Eelbacter' (genome Eelbacter_gp4_AA13) and '*Candidatus* Angelobacter' (genome Angelobacter_gp1_AA117), tentatively placed within the Blastocatellia and the Acidobacteriales, respectively. In the 7-Mb genome of *Candidatus* Eelbacter we identified 17 biosynthetic loci containing 74 NRPS and PKS open reading frames that were 404 kb in total length. In the 6.5-Mb genome of *Candidatus* Angelobacter there were 16 loci containing 54 NRP/PKS open reading frames that were 325 kb in total length. The biosynthetic genes from each species had only distant homology to those from the other. We confirmed the biosynthetic clusters for both genomes by re-analysing with 'Prediction Informatics for Secondary Metabolomes' (PRISM)[24] (Extended Data Figs. 2, 3). In total, each of these organisms contains over 900 kb of genes that are putatively involved in biosynthesis of secondary metabolites (about 12–14% of their recovered genomes). A phylogenetic analysis, using ribosomal protein sequences, of acidobacterial genomes from this study and reference databases revealed that both *Candidatus* Angelobacter

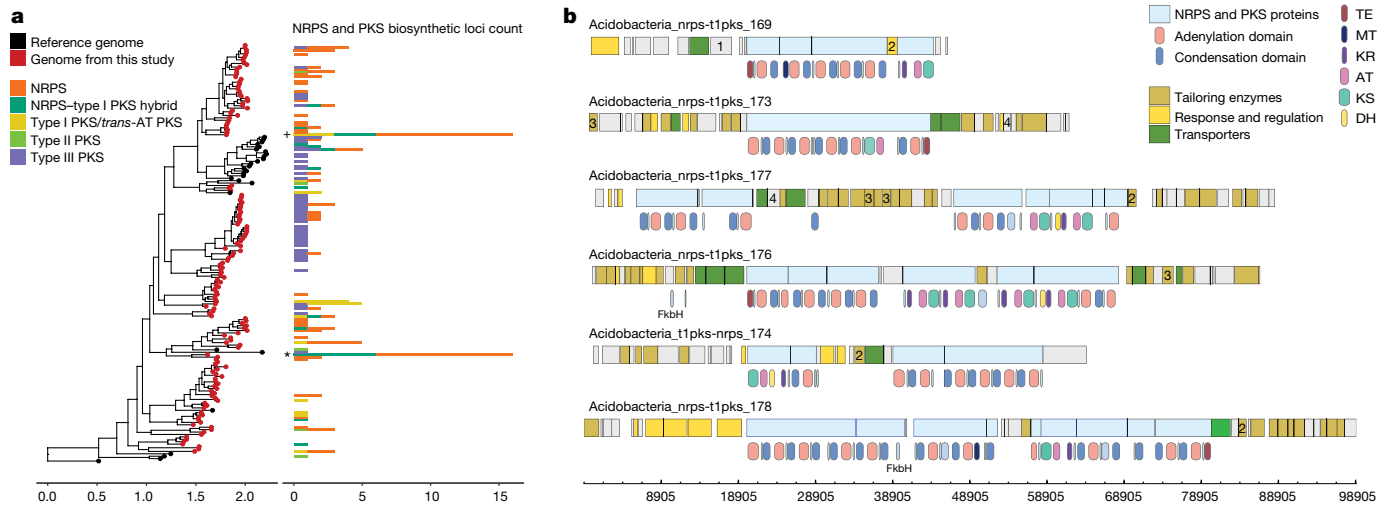


Fig. 2 | Biosynthetic NRPS and PKS loci from the Acidobacteria. **a**, Concatenated ribosomal protein phylogenetic tree of all acidobacterial genomes from this study (red) and existing reference genomes (black). Scale bar on the tree represents substitutions per site. Adjacent is a chart that reflects the count of NRPS and PKS biosynthetic gene clusters observed in each genome. The phylogenetic placements of *Candidatus* Eelbacter (*) and *Candidatus* Angelobacter (+) are marked. **b**, Six large PKS–NRPS hybrid biosynthesis gene clusters are encoded in the *Candidatus* Eelbacter genome. Predicted genes and biosynthetic protein domains are coloured by general function, and the genomic positions of polyketide and nonribosomal peptide synthetic domains are shown below each genome track. The following gene annotations are identified by number: 1, penicillin amidase; 2, oxygenase; 3, radical SAM proteins; and 4, betalactamase. AT, acyltransferase; DH, dehydrogenase; KR, ketoreductase; KS, ketosynthase; MT, methyltransferase; TE, thioesterase.

and *Candidatus* Eelbacter acquired their unusual arrays of biosynthetic operons independently in evolutionary time (Fig. 2a).

The *Candidatus* Angelobacter genomes included multiple lantibiotic biosynthesis proteins, a bacteriocin biosynthesis cluster, multigene operons with components for both a type VI and a type II secretion system, and several large RHS-repeat containing proteins, which have been hypothesized to have evolved to mediate microbial competition by facilitating transfer of protein toxins between species[25]. The *Candidatus* Eelbacter genome contained six clusters that were complex type I NRPS-PKS hybrid systems over 45 kb in length (Fig. 2b). Three replicate genomes of *Candidatus* Eelbacter were obtained from independent soil samples and shared the same set of biosynthetic clusters. Both species also possessed CRISPR–Cas loci (31 spacers and repeats in *Candidatus* Angelobacter and 438 across the *Candidatus* Eelbacter genome). The ecological and evolutionary forces that can select for the production of an unusually high number of metabolites in a species are varied, and previously characterized examples are microorganisms with complex cooperative lifestyles[26,27] or an association with a eukaryotic host[28]. The discovery of these two microorganisms establishes that bacterial specialization in secondary metabolite biosynthesis is not limited to known clades in the Actinomycetales, Proteobacteria, Cyanobacteria, Bacilli and the recently discovered Entotheonella[28]. When considered together, the genomic features of these Acidobacteria hint towards an unusually competitive lifestyle mediated by chemical and toxin production.

We tested whether the microorganisms genomically described in this study are active and express biosynthetic NRPS or PKS gene clusters by analysing metatranscriptomics data from 120 soil microcosm samples from two soil depths and two sampling locations from the same field site that were subject to amendment with glucose, methanol or water over 24 h (Supplementary Methods). These experiments were designed to probe the strong biological responses that occur in soils following water addition and nutrient release after a long dry period[29]. Because distinct NRPS or PKS clusters can produce products with very different bioactivities, we tracked expression of each gene cluster as a functional biosynthetic unit by pseudo-aligning exact matches of paired reads to full genomes obtained directly from the environment studied using Kallisto[30]. Overall, we detected expression for 198 NRPS and/or PKS genes across those NRPS and PKS clusters with any level of gene expression (133 out of 180 clusters) (Supplementary Table 8). Expression of NRPS and PKS clusters was detected in all four phyla that we studied, and 84 active clusters were detected in Acidobacteria (Extended Data Fig. 4). We detected the expression of genes within 10 biosynthetic clusters—including 11 genes with NRPS and/or PKS domains within these clusters—of *Candidatus* Eelbacter (Extended Data Fig. 5) and 14 clusters of *Candidatus* Angelobacter—including 25 genes with NRPS and/or PKS domains. We tested for co-expression of genes in all biosynthetic clusters and found that gene clusters were co-expressed more often than were randomized permutations of genes across each genome (Wilcoxon rank-sum test, $P < 0.001$).

Across all organisms in our dataset, we identified ten NRPS and/or PKS gene clusters from seven genomes with levels of expression that were time-dependent across the 24-h time course of the amendment experiments (permutational multivariate analysis of variance (PERMANOVA); $P < 0.05$, false discovery rate (FDR) = 5%) (Fig. 3a and Extended Data Fig. 6). We confirmed differential expression over time for individual genes within these clusters using a model that accounts for variation in both sequencing library sizes and organism abundances across samples[31] (DESeq2[32]; $P < 0.05$; FDR = 5%) (Supplementary Table 9). Notably, the expression of genes from several gene clusters in *Candidatus* Angelobacter showed a statistically significant increase 12–24 h after substrate addition (Fig. 3a), and we found that the expression of several biosynthetic genes of *Candidatus* Angelobacter was temporally distinct from the expression of core ribosomal genes (Fig. 3b). These results indicate that *Candidatus* Angelobacter populations respond to water and substrate addition, and independently regulate expression of secondary metabolite genes many hours after a period of increased core metabolic gene expression.

To predict the broader biological and ecological roles of these biosynthetic NRPS and PKS genes, we conducted separate co-expression analyses of all genes for each of the seven species identified with temporally dependent biosynthetic gene expression, using the WGCNA package[33] (Supplementary Methods), across the 120 microcosm timepoint samples. Co-expressed genes often share biological functions and regulation[34]. Modules of co-expressed genes significantly enriched in secondary metabolite genes were identified in four out of seven genomes ($P < 0.05$; hypergeometric distribution) (Fig. 3c, Extended Data Fig. 7 and Supplementary Table 10). These four modules were small (fewer than 69 genes) and very transcriptionally distinct. We found that all four secondary metabolism networks were dominated by genes involved in two-component systems, efflux and transcriptional

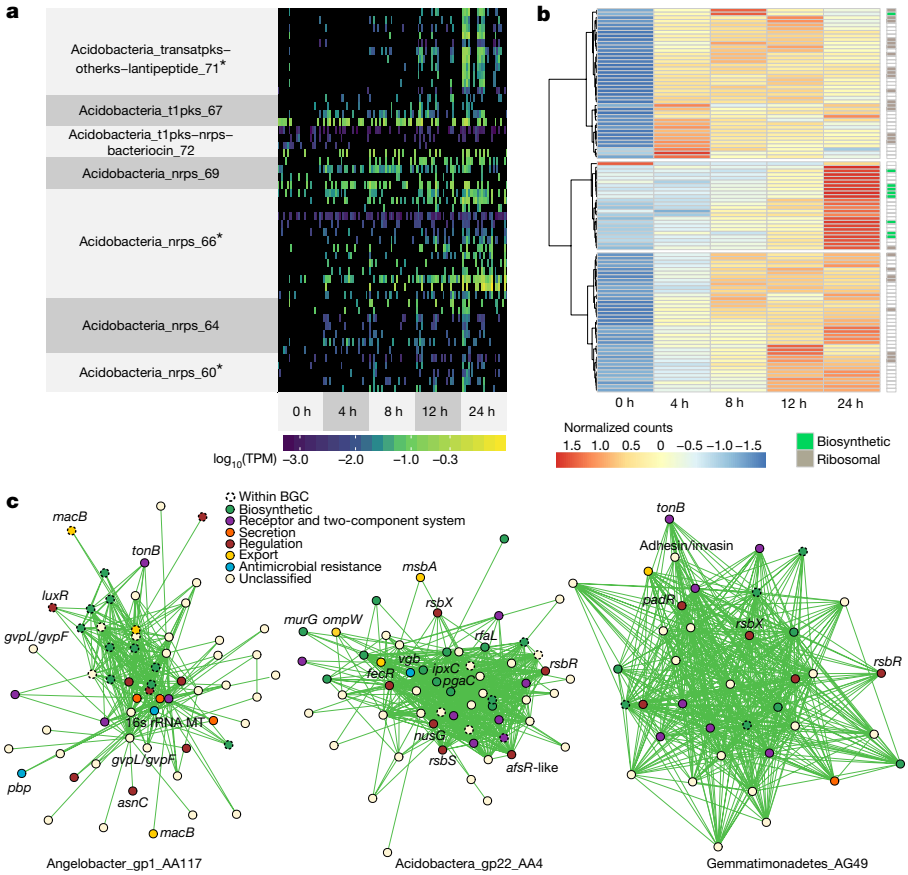


Fig. 3 | Metatranscriptomics of biosynthetic genes. a, Levels of transcriptional expression of genes from biosynthetic gene clusters encoded in the *Candidatus* Angelobacter genome, across 120 microcosm soil samples grouped by extraction times (reported in hours). Expression levels are reported in $\log_{10}$-transformed transcripts per million (TPM). Gene clusters that were significantly differentially expressed across time points (PERMANOVA); $^{*}P < 0.05$, FDR = 5% are marked by an asterisk. **b**, Hierarchical clustering of expression levels for differentially expressed ($n = 120$; DESeq2; $P < 0.05$; FDR = 5%) genes from the *Candidatus* Angelobacter genome across samples grouped by experimental time point. Differentially expressed genes from biosynthetic clusters and differentially expressed core ribosomal proteins are marked. Values are reported in counts transformed using the rlog transformation from DESeq2 and were normalized by row. **c**, The transcriptional co-expression network modules ($n = 120$ microcosm time-point samples) significantly enriched in NRPS and PKS biosynthetic genes from three genomes ($P < 0.05$; hypergeometric distribution). Nodes represent gene transcripts and edges between them represent high topological overlap values between the transcripts. Genes outlined are genes found within biosynthetic gene clusters (BGC), and are coloured by assigned function using the Kyoto Encyclopedia of Genes and Genomes and Pfam databases. 16s rRNA MT, gene encoding for a 16S rRNA methyltransferase.

regulators, and were almost completely devoid of genes for the core processes of transcription, translation and energy metabolism.

For *Candidatus* Angelobacter, genes from five biosynthetic clusters were co-expressed together in a module with a variety of genes involved in environmental sensing and response, including homologues of the gene that encodes for the iron siderophore uptake receptor TonB. Homologues of the gene that encodes for the macrolide export transporter MacB were also found to be co-expressed with the biosynthetic genes, as were two putative antimicrobial resistance genes—those encoding for penicillin-binding protein and for a 16S rRNA methyltransferase. Additional co-expressed genes included an operon for a type VI secretion system and an operon annotated as encoding for gas vesicle proteins. Notably, the Angelobacter population expressed biosynthetic genes from multiple clusters simultaneously, suggesting a concerted response that is linked to ecological competition.

Acidobacter_gp22_AA4 was found to co-express its NRPS gene cluster (Acidobacteria_nrps_112) with response-regulatory genes and a set of genes involved in cell surface structure remodelling, as well as an operon of genes involved in regulating stress response (*rsbX*, *rsbR* and *rsbS*). A homologue of virginiamycin B lyase (*vgb*), which is an inactivator of type B streptogramin antibiotics, was also co-expressed in this module. The same operon of genes involved in the regulation of stress response was found to be co-expressed in the transcriptional network containing a biosynthetic cluster (cluster Gemmatimonadetes_nrps_183) in Gemmatimonadetes_AG49, along with a *tonB* homologue.

In summary, we uncovered extensive evidence for secondary metabolite synthesis in a large collection of bacterial genomes from four phyla of soil bacteria that have not previously been genomically linked to this capacity. Although we cannot confidently predict more than the basic chemical scaffolds of the products derived from the biosynthetic genes reported here, or their biological activities, a large percentage of known polyketide and nonribosomal metabolites isolated from microbial sources have antimicrobial activity[35]. Transcriptional associations between specific NRPS and PKS gene clusters, regulators of iron metabolism and putative antimicrobial resistance mechanisms suggest that these gene clusters may be involved in competition for iron resources and antibiotic production. The findings underline the utility of genome-resolved metagenomic investigations of soil ecosystems and open the way for laboratory characterization of genes for novel bioactive metabolites with potential ecological and pharmaceutical importance.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0207-y.

Received: 2 August 2017; Accepted: 2 May 2018;
Published online 13 June 2018.

1. Hibbing, M. E., Fuqua, C., Parsek, M. R. & Brook Peterson, S. Bacterial competition: surviving and thriving in the microbial jungle. *Nat. Rev. Microbiol.* **8**, 15–25 (2010).
2. Charlop-Powers, Z., Owen, J. G., Reddy, B. V., Ternei, M. A. & Brady, S. F. Chemical–biogeographic survey of secondary metabolism in soil. *Proc. Natl Acad. Sci. USA* **111**, 3757–3762 (2014).
3. Cragg, G. M. & Newman, D. J. Natural products: a continuing source of novel drug leads. *Biochim. Biophys. Acta* **1830**, 3670–3695 (2013).
4. Rappé, M. S. & Giovannoni, S. J. The uncultured microbial majority. *Annu. Rev. Microbiol.* **57**, 369–394 (2003).
5. Fierer, N. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat. Rev. Microbiol.* **15**, 579–590 (2017).
6. Bergmann, G. T. et al. The under-recognized dominance of Verrucomicrobia in soil bacterial communities. *Soil Biol. Biochem.* **43**, 1450–1455 (2011).
7. Kielak, A. M., Barreto, C. C., Kowalchuk, G. A., van Veen, J. A. & Kuramae, E. E. The ecology of Acidobacteria: moving beyond genes and genomes. *Front. Microbiol.* **7**, 744 (2016).
8. Butterfield, C. N. et al. Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone. *PeerJ* **4**, e2687 (2016).
9. DeBruyn, J. M., Nixon, L. T., Fawaz, M. N., Johnson, A. M. & Radosevich, M. Global biogeography and quantitative seasonal dynamics of Gemmatimonadetes in soil. *Appl. Environ. Microbiol.* **77**, 6295–6300 (2011).
10. Weber, T. et al. antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* **43**, W237–W243 (2015).
11. Medema, M. H. et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* **39**, W339–W346 (2011).
12. Hadjithomas, M. et al. IMG-ABC: a knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. *MBio* **6**, e00932-e15 (2015).
13. Cimermancic, P. et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421 (2014).
14. Wang, H., Fewer, D. P., Holm, L., Rouhiainen, L. & Sivonen, K. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. *Proc. Natl Acad. Sci. USA* **111**, 9259–9264 (2014).
15. Parsley, L. C. et al. Polyketide synthase pathways identified from a metagenomic library are derived from soil Acidobacteria. *FEMS Microbiol. Ecol.* **78**, 176–187 (2011).
16. Rondon, M. R. et al. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* **66**, 2541–2547 (2000).
17. Charlop-Powers, Z. et al. Global biogeographic sampling of bacterial secondary metabolism. *eLife* **4**, e05048 (2015).
18. Fischbach, M. A. & Walsh, C. T. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. *Chem. Rev.* **106**, 3468–3496 (2006).
19. Medema, M. H. et al. Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.* **11**, 625–631 (2015).
20. Medema, M. H., et al. A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. *PLoS Comput. Biol.* **10**, e1004016 (2014).
21. Thaker, M. N. et al. Identifying producers of antibacterial compounds by screening for antibiotic resistance. *Nat. Biotechnol.* **31**, 922–927 (2013).
22. Johnston, C. W. et al. Assembly and clustering of natural antibiotics guides target identification. *Nat. Chem. Biol.* **12**, 233–239 (2016).
23. Gibson, M. K., Forsberg, K. J. & Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* **9**, 207–216 (2015).
24. Skinnider, M. A., Merwin, N. J., Johnston, C. W. & Magarvey, N. A. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res.* **45**, W49–W54 (2017).
25. Koskiniemi, S. et al. Rhs proteins from diverse bacteria mediate intercellular competition. *Proc. Natl Acad. Sci. USA* **110**, 7032–7037 (2013).
26. Claessen, D., de Jong, W., Dijkhuizen, L. & Wösten, H. A. Regulation of *Streptomyces* development: reach for the sky. *Trends Microbiol.* **14**, 313–319 (2006).
27. Zhang, Y., Ducret, A., Shaevitz, J. & Mignot, T. From individual cell motility to collective behaviors: insights from a prokaryote, *Myxococcus xanthus*. *FEMS Microbiol. Rev.* **36**, 149–164 (2012).
28. Wilson, M. C. et al. An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **506**, 58–62 (2014).
29. Unger, S. et al. The influence of precipitation pulses on soil respiration–assessing the "Birch effect" by stable carbon isotopes. *Soil Biol. Biochem.* **42**, 1800–1810 (2010).
30. Bray, N., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
31. Klingenberg, H. & Meinicke, P. How to normalize metatranscriptomic count data for differential expression analysis. *PeerJ* **5**, e3859 (2017).
32. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
33. Langfelder, P & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
34. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
35. Bérdy, J. Bioactive microbial metabolites. *J. Antibiot. (Tokyo)* **58**, 1–26 (2005).

Acknowledgements We thank S. Spaulding for assistance with fieldwork, and M. Traxler and W. Zhang for helpful discussions. Sequencing was carried out under a Community Sequencing Project at the Joint Genome Institute. Funding was provided by the Office of Science, Office of Biological and Environmental Research, of the US Department of Energy Grant DOE-SC10010566, the Paul G. Allen Family Foundation and the Innovative Genomics Institute of the University of California, Berkeley.

Author contributions A.C.-C. performed genomic and transcriptomic analysis; S.D. performed metagenome assembly and curation; C.N.B. performed microcosm experiments and RNA extractions; A.C.-C., S.D. and J.F.B. wrote the manuscript; B.C.T. supported the metagenomics bioinformatics work; and J.F.B. supervised the project.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0207-y.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0207-y.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.F.B.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessments.

Soil sampling and DNA extraction. Soil samples were collected from the Angelo Coast Range Reserve meadow (39° 44′ 21.4″ N 123° 37′ 51.0″ W) on four dates in 2014 that bracketed the first winter rain of the season. Samples were collected from three depths, 10–20 cm, 20–30 cm and 30–40 cm at six independent sampling sites that were first metagenomically characterized as part of a previous study[8]. Sampling was conducted in biological triplicate, with three of the sites being unamended biological control plots and three being amended with extended spring rainfall from a sprinkler system as described in a previous publication[8]. Sampling was accomplished using a soil coring device that was fitted with sterilized polycarbonate sheaths. Sheaths were removed after each collection event. After collection, samples were flash-frozen in a mixture of dry ice and ethanol, and placed on dry ice for transport. A total of 60 soil cores were sampled across all depth and treatment conditions.

For each depth, DNA was extracted using MoBio Laboratories PowerMax Soil DNA Isolation kits from 10 g of soil as previously described[8]. Mean DNA concentration in the extracted samples, quantified by using qubit fluorometric assay, was 388 ng/μl.

Sequencing, genomic assembly and binning. Metagenomic libraries for all 60 samples were prepared and sequenced at the Joint Genome Institute using an Illumina HiSeq 2500 platform to generate 250-bp paired-end reads. Samples were multiplexed for sequencing. Raw sequence data were processed with BBmap[36] to remove Illumina adaptor and phiX sequences, and reads were quality-score trimmed using Sickle with default parameters[37]. Read sets were subsequently analysed for per-base GC content using FastQC[38], and it was determined that GC content increased substantially after 200 bp in some sample read sets. Thus all reads longer than 200 bp were hard-trimmed to 200 bp using BBmap. In total, 6.22×10^9 reads were sequenced across all samples, which yielded 1.24 Tb of total sequence information with an average read count of 1.04×10^8 reads per sample.

The 60 samples were individually assembled de novo on a 24-core Intel Xenon Linux cluster node with 256 Gb of RAM using IDBA-UD[39] with the following initial parameters: –pre_correction,–mink 30,–maxk 200,–step 10. In the 13 cases in which assemblies did not complete owing to memory requirements, minimum *k*-mer size was increased to 40 bp. The resulting assemblies averaged 1.15 Gb of assembled sequence with an N50 of 1,609 bp. Sequencing coverage of each contig was calculated by mapping raw reads back to assemblies using Bowtie2[40]; 36.4% of reads mapped back to assembled sequence on average. It should also be noted that contigs >100 kb in length were acquired from all 60 assemblies, with a maximum contig size across assemblies of 2.7 Mb.

All resulting assemblies were subsequently clustered into genome bins individually using a hybrid binning approach. Initially, reads from all assemblies were separately cross-mapped to all scaffolds >2 kb in size from a single assembly using Bowtie2 to generate a coverage profile for the scaffolds of that assembly across all samples. Scaffold differential coverage profiles were used to inform five separate automated binning software packages: ABAWCA, ABAWACA2[41], MaxBin2[42], CONCOCT[43] and MetaBAT[44], which were run on all samples individually. The resulting output genome bins for all packages run on a single sample were combined, assessed for completeness using an inventory of 51 universal single-copy genes (SCGs), and dereplicated by selecting the most complete bin of an overlapping set using DAStool[45]. Following automated binning, all genomic bins were manually inspected and curated using our in-house bin visualization and analysis system, ggKbase[46] (http://ggkbase.berkeley.edu). Finally, after manual curation in ggKbase, reads from a given sample were mapped back to the bins derived from that sample to identify and correct assembly and scaffolding errors, as previously described[47]. In total, 10,463 individual genome bins were identified across all samples. Of these bins, 3,334 were then estimated at a completeness of ≥ 70% using CheckM[48]. Taxonomic assignment of bins was performed by looking at the closest known hits and phylogenetic placement of ribosomal marker proteins. Bins were then dereplicated by clustering their ribosomal S3 proteins at 99% amino acid identity and choosing the bin in each cluster with the highest completeness and lowest contamination, which resulted in a final set of 377 nonredundant bins in the bacterial phyla of interest.

Genomic analysis of genomes and biosynthetic gene clusters. Curated genomes were individually processed using antiSMASH 3.0[10] with default parameters. The results are summarized in Supplementary Table 2 for gene clusters on contigs greater than 10 kb, Supplementary Table 1 for gene clusters on contigs smaller than 10 kb and Supplementary Table 4 for all PKS and NRPS clusters on contigs greater than 10 kb. Ribosomal protein phylogenetic trees were built using a concatenated set of 16 ribosomal proteins[49] for all Acidobacteria genomes in this dataset, as well as those that could be obtained from GenBank or the Integrated Microbial Genomes platform. An *Escherichia coli* genome was used as an outgroup for the tree. These protein sequences were aligned with MUSCLE[50] and then a maximum likelihood phylogeny was built using FastTree2[51] with default parameters.

To test whether existing primer-based methods have the ability to amplify these biosynthetic gene sequences, sets of forward and reverse degenerate primers used by previous analyses of biosynthetic genetic diversity[2,17] for ketosynthase genes and adenylation domain genes were searched for pattern matches against all NRPS and PKS clusters in both reverse and forward reading frames. The inosine nucleotides were substituted with the ambiguous code B, because these nucleotides can base pair with adenine, cytosine and uracil. Only five of our gene clusters had correctly oriented matches to both a forward and reverse primer within 2 kb of each other (Supplementary Table 6).

The network of gene clusters based on shared gene content was built by performing an all-versus-all BLASTP search of predicted biosynthetic protein sequences. Shared proteins were defined as protein alignments with at least 50% of the query sequence covered and amino acid per cent identity >50%. Two clusters (nodes) were connected if either one shared at least 10% of its proteins with the other. The width and colour intensity of the network edges was scaled with the length of the shared protein alignments, normalized to the length in base pairs of the two clusters being compared. Biosynthetic gene clusters were compared to clusters previously reported in the MiBIG repository[19] using BLASTP and the same definition of shared proteins, and the closest hits to MiBIG clusters containing at least five genes were reported. To identify antibiotic resistance genes in clusters, we searched protein products of all biosynthetic gene clusters with a set of hidden Markov models derived from a previous publication[22], using HMMER with the gathering threshold cutoffs specified in this previous study. We then manually curated hits and eliminated matches to ambiguous functions (acetyltransferases, general methyltransferases and amidases) and focused on reporting proteins with functions that are unlikely to be involved in generic biosynthetic pathways. The *Candidatus* Angelobacter and Eelbacter genomes were both subsequently analysed using the PRISM3 webserver[24].

Soil microcosm experiments and RNA extraction. At the Angelo Coast Range Reserve meadow, five holes were bored within a 1-m^2 area to obtain 10-cm-long cores of soil, from depths 10–20 cm and 30–40 cm (permission under APP # 27790). Samples were collected on 21 September 2015. At each depth, five cores were mixed in a large Whirl-Pak bag, then distributed into five capped core liners and stored in individual Whirl-Pak bags at 4 °C. The unsieved soils were mixed a second time in the laboratory to obtain six equally proportioned samples, and the weights were measured. To settle the soil, the core liners were struck with a rubber mallet 50 times each, and then stored at 4 °C. The night before wet-up experiments, the cores were placed in a cooler alongside the substrate that was to be added, so that the soil and substrate equilibrated to the same temperature and the soil would be kept in the dark. Immediately before adding the substrate, 10 g soil was collected for DNA extraction and 2 g soil with 4 ml LifeGuard RNA Soil Preservation Solution (MoBio) was collected for RNA purification. Both were immediately frozen in liquid N_2 and stored in a freezer at −80 °C. Samples at different time points were collected for nucleic acid extraction in the same manner. Ten millimolar glucose, methanol or water substrate was added to the open-soil core liners and soil in a cooler by pipette 2.5–4 ml at a time over 1 min, and the lid was closed. Substrates were added in amounts that increased the soil moisture to the level of a sample collected from the meadow after 29 cm of rainfall on 5 November 2015 (the moisture level of the field sample was determined by weight loss on drying). RNA was isolated from 2 g soil with RNA PowerSoil Total RNA Isolation kits, following kit protocols. cDNA libraries were prepared and were sequenced to generate 5.9×10^9 150-bp paired-end reads.

Transcriptomics. To test for the expression of clusters of biosynthetic genes within a soil environment, we analysed metatranscriptomics data from experimental soil microcosms. Soil samples from depths of 20 cm and 40 cm from two sampling locations were subject to amendment with glucose, methanol or water, and RNA was extracted from samples at 0, 4, 8, 12 and 24 h after treatment. From the 120 sequenced samples, we generated 5.9×10^9 150-bp paired-end reads. Transcript abundances for all Prodigal-predicted gene sequences from all genomes reconstructed from the project site were quantified using Kallisto[30] exact pseudoalignments of paired reads. Kallisto was run using default parameters. Transcripts that were either found to be expressed in at least 10% of samples or to have at least 100 counts were reported and included in downstream analyses. Differential gene expression analysis was performed using PERMANOVA and DESeq2[32] (see 'Statistical analysis').

We mapped RNA reads from one replicate for each sample at the $t = 0$ and $t = 24$ h time points to 16S sequences assembled from our genomic data from the two plots from which the microcosm soil was obtained. A subset of 4,000 RNA reads was compared to the SILVA 16S database using BLAST to determine the percentage of RNA reads that were 16S rRNAs. Of 16S rRNA reads in the RNA data, 47% ± 19% were determined to be at least 98% identical to 16S sequences assembled in the genomic data (Supplementary Table 12), which indicates that the

community that we assembled in the genomic dataset is a substantial fraction of the active community in the metatranscriptomic data.

We performed weighted gene co-expression network analyses using the WGCNA package[33] separately and individually on genes from seven genomes that were identified as having differentially expressed biosynthetic gene clusters over time, reasoning that these genomes will have the strongest signal of secondary metabolite co-expression. Transcripts per million for each gene were log-transformed. A soft network threshold was generated by choosing the lowest value that returned an R^2 fit to a scale-free network greater than 0.8. A signed adjacency matrix was built using Pearson correlations, and a topographical overlap matrix was generated from the adjacency matrix. Module detection was run using the cuttreeDynamic() function with the 'hybrid' method, a minimum cluster size of 15, deepSplit set to TRUE and a cutHeight of 0.95.

Statistical analysis. To test whether cluster genes were significantly more co-expressed than random genes across a genome, we calculated all Spearman correlations between genes within clusters (mean $\rho = 0.063$; $n = 5{,}940$ comparisons), and compared this distribution of correlations to a distribution of all Spearman correlations between 100 randomly chosen genes from each genome (mean $\rho = 0.041$; $n = 503{,}699$ comparisons) using an independent two-group Wilcoxon rank-sum test ($P < 0.001$). We also compared both distributions to a distribution of randomly selected genes from the entire dataset compared (mean $\rho = 0.026$ $n = 4947228$ comparisons) and found random genes to have the lowest levels of co-expression ($P < 0.001$).

To identify differentially expressed clusters of genes between time points, we used the adonis function from the vegan package[52]. Transcript abundances in transcripts per million were $\log_2$-transformed, and adonis tests were run on all clusters with any expression data for at least five proteins. *P* values were corrected for multiple tests using the Benjamini and Hochberg[53] method with a controlled family wise error rate of 5%.

To detect differential expression of individual genes within differentially expressed biosynthetic clusters between time points, we modelled Kallisto counts in the context of all metadata variables (plot, depth, treatment and time) using a negative binomial model implemented in DESeq2[32]. Kallisto count data from each genome were analysed independently so that the DEseq size factors for cross-sample count normalization would reflect the total transcriptomic activity of that genome in each sample. This approach is robust to biases in total transcriptomic activity per organism between samples, with the intention to identify differences in gene expression independent of changes in taxonomic composition, similar to previously reported methods[30]. After size factor normalization, counts were fit to a negative binomial model of the form: count ~ depth + plot + treatment + time. To specifically test whether any genes exhibit differential expression associated with changes in time while accounting for the effects of depth, plot and treatment, we fit count data to a reduced model of the form: count ~ depth + plot + treatment. We then compared fits between the full and reduced model using the likelihood ratio test implemented in DESeq2. The significant genes (with an FDR-corrected $P < 0.05$) identified by comparing the full and reduced model were grouped, and direct comparisons were made between counts at 0 h and all other time points, to find those time points that exhibited a significant change in expression relative to the 0 h time point. This method confirmed differential expression of several individual genes within each differentially expressed biosynthetic cluster.

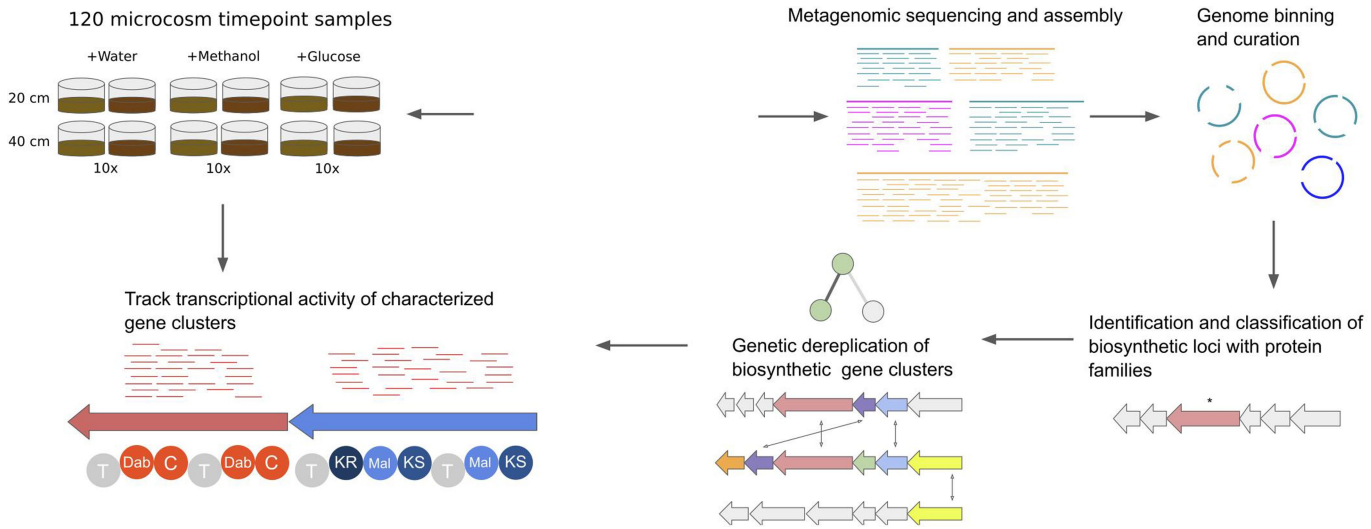
When examining modules of co-expression genes, the hypergeometric test was used to determine whether a module was significantly enriched in biosynthetic genes, using the phyper function in R.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. Custom code used for the analyses (transcriptomics expression, DESeq2 differential expression and WGCNA co-expression analyses) that support this work is available in R Notebook format at http://www.github.com/alexcritschristoph/angelo_biosynthetic_genes_analysis.

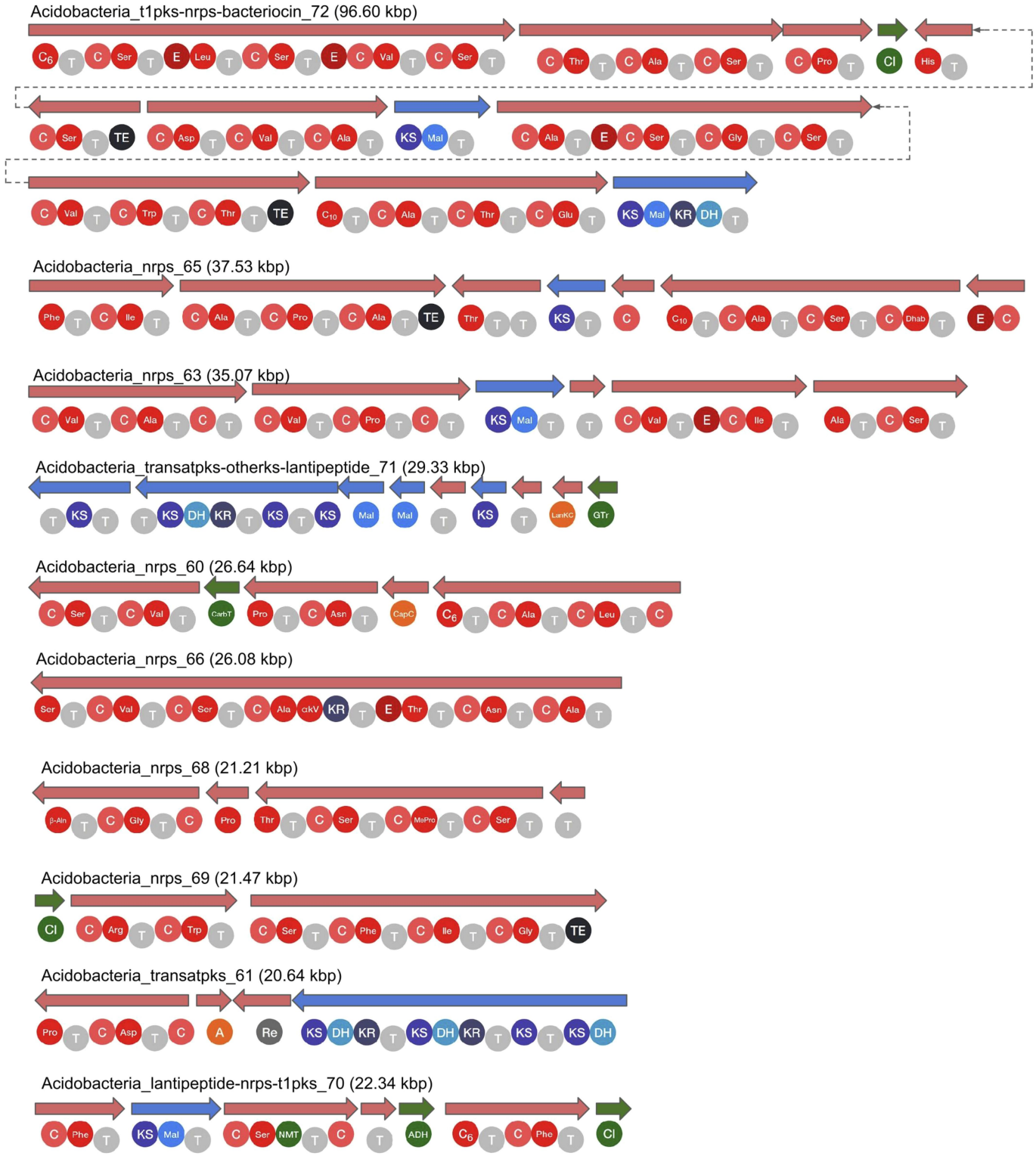
Data availability. All genomic data associated with this project has been deposited in BioProject under accession PRJNA449266. DNA sequencing reads for this project have been deposited in the Sequence Read Archive database under PRJNA449266. Genomes analysed as part of this project have been submitted to the Whole Genome Shotgun (WGS) database. Genomes are also available through ggKBase at the following URL: http://ggkbase.berkeley.edu/angelo2014/organisms. Raw data for Fig. 2a and AntiSMASH annotated GenBank files for biosynthetic gene clusters reported on in this Letter are available at: http://www.github.com/alexcritschristoph/angelo_biosynthetic_genes_analysis.

36. Bushnell, B. BBMap short read aligner. http://sourceforge.net/projects/bbmap (University of California, Berkeley, 2016).
37. Joshi, N. A. & Fass, J. N. sickle - a windowed adapative trimming tool for FastQ files (version 1.33) https://github.com/najoshi/sickle (2011).
38. Andrews, S. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).
39. Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
40. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
41. Brown, C.T. et al. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).
42. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
43. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
44. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
45. Sieber, C. M. K. et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Methods* https://doi.org/10.1038/s41564-018-0171-1 (2018).
46. Banfield, J. *Development of a Knowledgebase to Integrate, Analyze, Distribute, and Visualize Microbial Community Systems Biology Data.* Report No. DOE-UCB-4918) (US Department of Energy, 2015).
47. Anantharaman, K. et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* **7**, 13219 (2016).
48. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
49. Hug, L. A. et al. A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
50. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
51. Price, M. N., Dehal, P. S. and Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS ONE* **5**, e9490 (2010).
52. Oksanen, J. et al. vegan: Community ecology package https://cran.r-project.org/package=vegan (2007).
53. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).

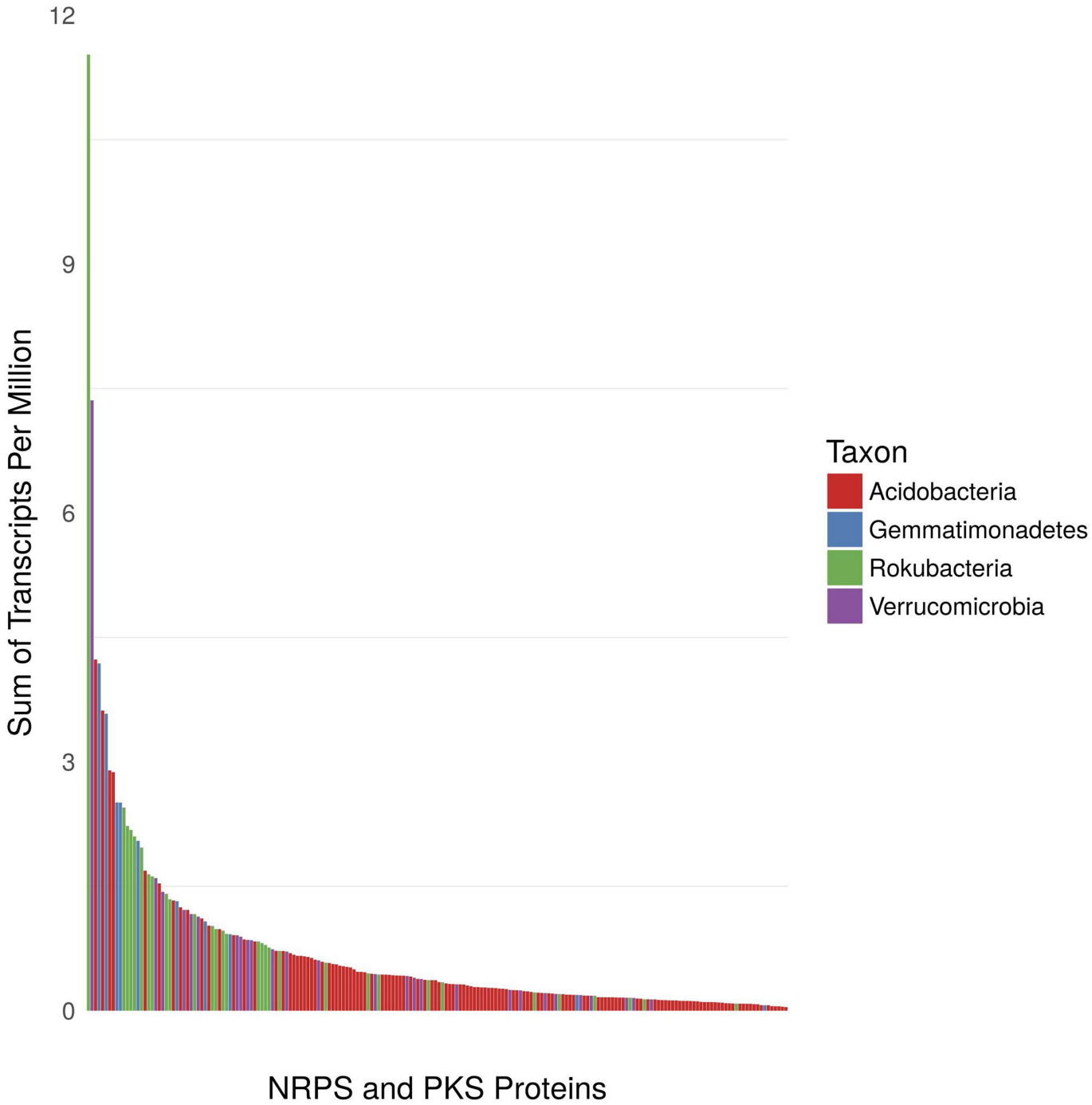


Extended Data Fig. 1 | Experimental plan and project overview. Schematic showing major components of microcosm time-point sampling and metagenomic analyses.

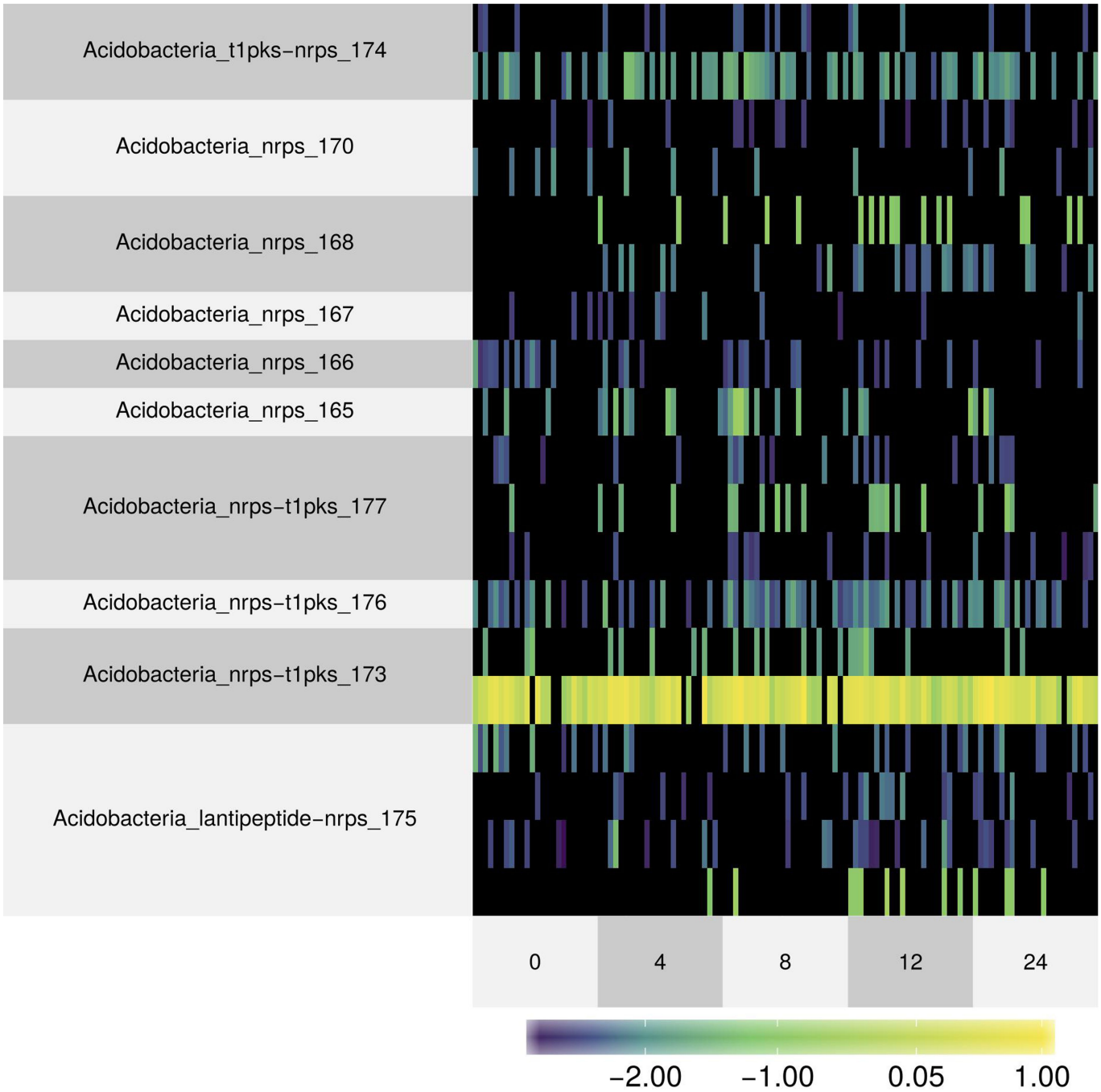
Extended Data Fig. 2 | NRPS and PKS biosynthetic loci of the *Candidatus* Eelbacter genome. Biosynthetic loci identified by both antiSMASH and PRISM from the *Candidatus* Eelbacter genome that contained at least 10 kb of biosynthetic genes. Predictions of the organization of the biosynthetic domains in each locus shown here were determined by PRISM. Smaller biosynthetic loci from this genome are not shown. Full names for the biosynthetic domains are given in Supplementary Table 11.



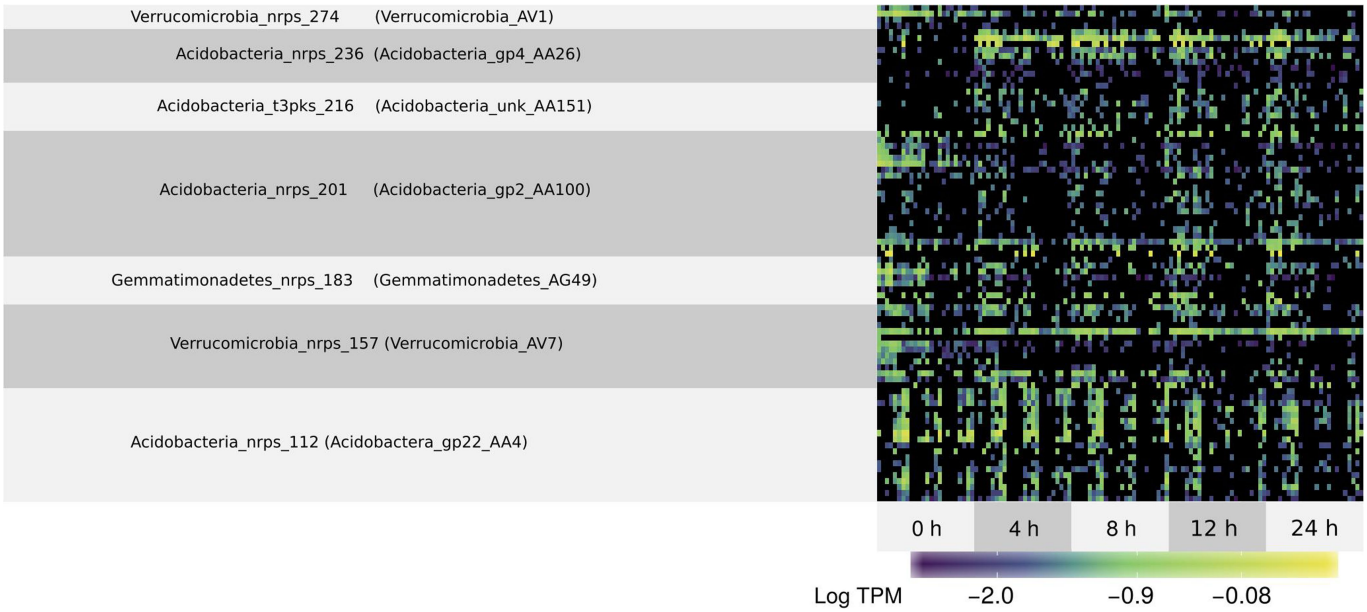
Extended Data Fig. 3 | NRPS and PKS biosynthetic loci of the *Candidatus* Angelobacter genome. Biosynthetic loci identified by both antiSMASH and PRISM from the *Candidatus* Angelobacter genome that contained at least 10 kb of biosynthetic genes. Predictions of the organization of the biosynthetic domains in each locus shown here were determined by PRISM. Smaller biosynthetic loci from this genome are not shown. Full names for the biosynthetic domains are given in Supplementary Table 11.



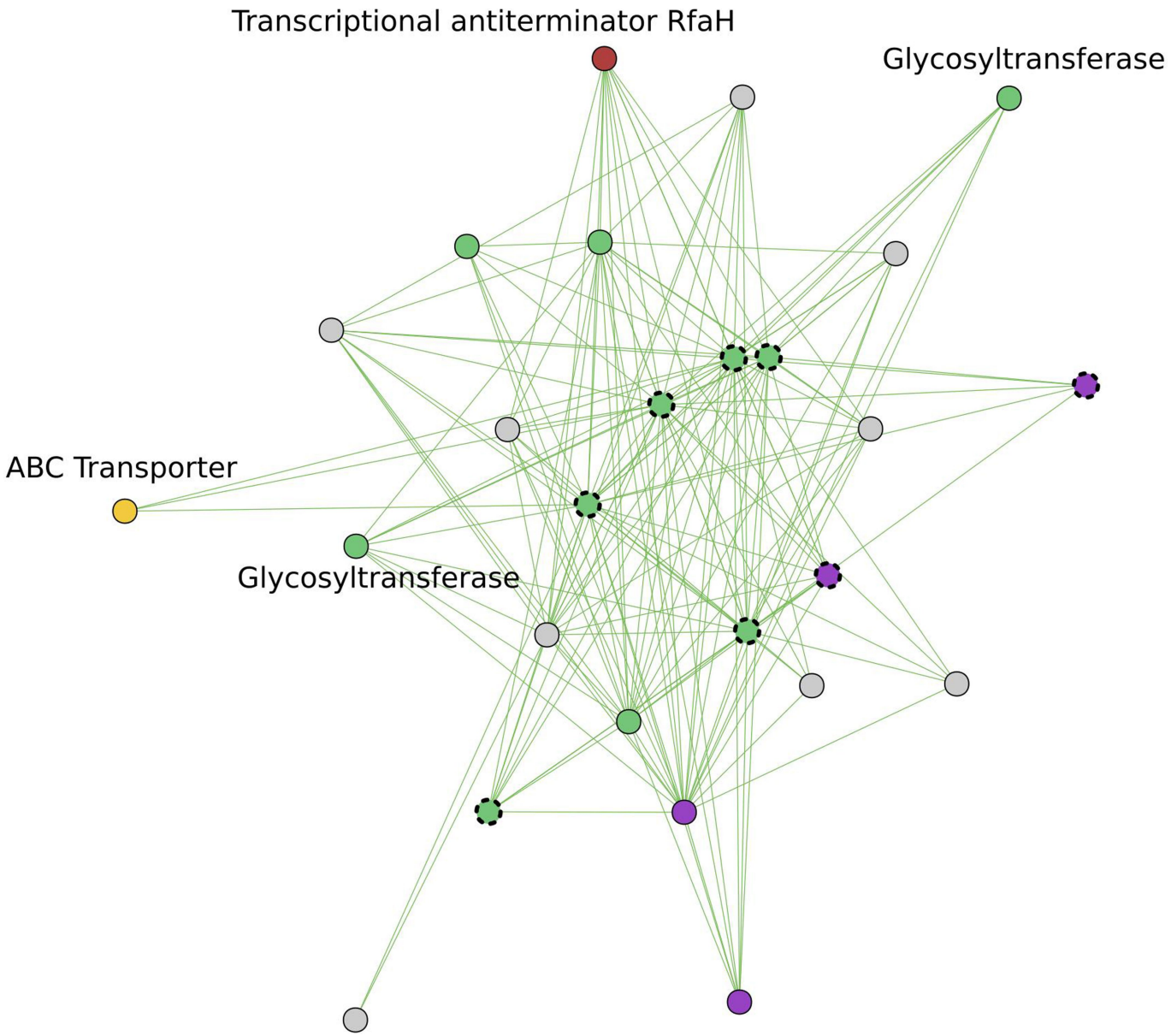
Extended Data Fig. 4 | Metatranscriptomics of NRPS and PKS proteins. The graph shows levels of transcriptional expression of genes containing NRPS and PKS protein domains across genomes from the four phyla of interest. Values are reported in $\log_{10}$-transformed transcripts per million and are summed across the 120 soil microcosm samples.



Extended Data Fig. 5 | Metatranscriptomics of the *Candidatus* Eelbacter genome. The levels of transcriptional expression of genes from biosynthetic gene clusters encoded in the *Candidatus* Eelbacter genome across 120 soil microcosm time-point samples grouped by extraction times (reported in hours) are shown. Expression levels are reported in $\log_{10}$-transformed transcripts per million.



Extended Data Fig. 6 | Differentially expressed biosynthetic gene clusters over time. The levels of expression of biosynthetic gene clusters from all organisms studied (excluding *Candidatus* Angelobacter data shown in Fig. 3a) that were found to be significantly differentially expressed between time points (PERMANOVA; $n = 120$; $P < 0.05$, FDR = 5%) across 120 soil microcosm time-point samples are shown. Expression levels are reported in $\log_{10}$ transcripts per million.



Extended Data Fig. 7 | Biosynthetic co-expression transcriptional module from Verrucomicrobia_AV7. A transcriptional network of co-expressed Verrucomicrobia_AV7 genes from a module found to be significantly enriched in genes from the biosynthetic gene clusters Verrucomicrobia_nrps_156 and Verrucomicrobia_nrps_157 ($P < 0.05$; hypergeometric distribution) is shown. Genes from the biosynthetic locus are outlined with a dashed line.

nature research

Corresponding author(s): Jillian Banfield

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software used for data collection, which was limited to sequencing data.

Data analysis	Data analysis: Kallisto 0.43 WGCNA 1.63 DESeq2 1.18 Sickle 1.33 IDBA_UD 1.1.0 Bowtie2 2.2.6 CONCOCT 0.4 MetaBAT 2 DAS Tool 1.1 CheckM 1.0.10 MUSCLE v3.8.31 FastTree 2.1 Antismash 3.0.5.1 Workflows describing the custom analysis code used in this study are available as described in the Code Availability Statement.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All genomic data associated with this project has been deposited and linked to the BioProject PRJNA449266. DNA sequencing reads for this project have been deposited in the SRA database under that BioProject number. Genomes analyzed as part of this project were submitted to the Assembly, WGS, and NR databases. Genomes are also available through ggKBase at the following URL: ggkbase.berkeley.edu/angelo2014/organisms. Raw data for Figures 1-3 and AntiSMASH annotated Genbank files for biosynthetic gene clusters reported on in this study are available at: github.com/alexcritschristoph/angelo_biosynthetic_genes_analysis

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size used for the metatranscriptomics experiment was 120 samples, with a structured design of 3 substrate amendment conditions, soil from 2 depths and 2 different plots, across 5 time points. The n=120 sample size stems from the sample size of each homogenous block of soil (homogenous plot, depth, and substrate) having 10 samples (2 per time point). A formal analysis of statistical power was not performed, but this sample size was chosen based on an evaluation of sample sizes for microbial transcriptomics experiments in existing literature, and made even larger to compare signals across different sample groupings.
Data exclusions	No data were excluded beyond sequencing reads with low quality scores, transcriptomics sequencing reads that did not map to genes within our dataset, and mapped transcripts with extremely low abundance, as is commonly performed in transcriptomics experiments.
Replication	One of our primary findings, the reconstruction of a genome encoding for an unusual number of biosynthetic gene clusters, was replicated across independent assembly and binning pipelines in three different samples. Further replication of the experiment was not attempted.
Randomization	No randomization was performed as part of this study.
Blinding	No blinding was performed during DNA and RNA extractions in this study. It is unlikely that the investigators would have been able to influence sequencing results in a systematic manner during the sample extraction and processing.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging