LAWRENCE LIVERMORE NATIONAL LABORATORY

LLNL-TR-818366

BSSD Q1 2021 Performance Metric report: LLNL Soil Microbiome SFA

J. Pett-Ridge

January 14, 2021

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

LLNL-TR-818366: BSSD 2021 Performance Metric Q1

Goal: Develop new omics-based techniques to understand microbiome function in environmental samples

Q1 Target: Report on the latest genomic-based techniques used to explore the composition of microbiomes in environmental samples.

Introduction

The LLNL "Microbes Persist" Soil Microbiome Scientific Focus Area (SFA) seeks to determine how microbial soil ecophysiology, population dynamics, and microbe-mineral-organic matter interactions regulate the persistence of microbial residues and the formation of soil carbon. Our SFA research program is now four years old; it evolved and benefited from previously-funded BSSD projects in the Firestone (UCB), Banfield (UCB), Sullivan (OSU) and Hungate (NAU) labs. We use stable isotope probing in combination with 'omics to measure how changing water regimes shape activity of individual microbial populations and ecophysiological traits that affect the fate of microbial and plant C. Using measures of population dynamics and microbiome-mineral interactions, we are working to synthesize both genome-scale and ecosystem-scale models of soil organic matter (SOM) turnover, to predict the long-aspired connection between soil microbiomes and fate of soil C.

One of the critical first steps in microbiome analysis is to simply understand 'who's there'. In the past decade, our microbiome community characterization efforts have moved from compositional analyses based on gene-based amplicon sequencing (454 pyrosequencing and later high-throughput Illumina based) targeting 16S/18S rRNA genes and ITS[1, 2] and functional gene arrays (GeoChip)[3], to more comprehensive approaches such as metagenomics, viromics and metatranscriptomics—where total soil DNA, viral fraction DNA, and mRNA are shotgun sequenced, and reconstructed into large contigs or near-complete genome assemblies (metagenome-assembled-genomes or MAGs) for individual populations. At the same time, we have begun to gather **multi-domain composition information** in novel ways, using barcoding approaches to target protists by amplifying 18S rRNA genes, or direct rRNA and DNA shotgun sequencing to simultaneously study Bacteria, Archaea, and Eukarya populations without amplification. We also use co-occurrence network and community assembly analyses to discern the likelihood of cross-kingdom interactions and the ecological relationships between microbiome taxa. Due to our growing interest in the **role of soil viruses**, we are exploring links between viruses and hosts in both DNA and RNA datasets, and sequences in viral auxiliary metabolic genes (AMGs) that may convey functional capabilities. In all our datasets, we aim for a high level of sample and temporal replication, and frequently harness the power of **stable isotope probing** (SIP) which allows us to focus on the active taxa in a microbiome[4-9].

I. Whole Microbial Community Analysis

Characterizing 'relic' DNA in soil: Microbes exist in different metabolic states in soils, with differing degrees of influence on the environment. However, traditional DNA extraction and sequencing approaches do not distinguish taxa that are active from inactive, nor live from dead. There is increasing recognition that a substantial proportion of soil DNA may be 'relic DNA', extracellular (non-viable) DNA from dead microorganisms that can persist in soil for long periods of time. This relic DNA can fundamentally compromise interpretations of soil microbiome community composition. Our colleagues in the Fierer Lab at the University of Colorado, Boulder have showed that when unaccounted for, this relic DNA pool can have significant effects on estimates of soil microbial community abundance and composition[10]. While there are several methods for estimating or removing relic DNA from mixed community samples, these approaches are not well-tested and do not provide information about microbial activity. To differentiate between actively growing microorganisms, dead and degraded DNA, and dead and stabilized extracellular DNA, we are working to combine measurements of relic DNA with heavy water stable isotope probing (SIP)[1, 11]. We are testing an established method[10] to characterize relic DNA in soil that uses a photoreactive, intercalating dye, propidium monoazide (PMA). PMA cross-links with DNA, rendering it unamplifiable via PCR and unsequenceable. Thus, by quantifying gene abundance from soils treated with/without PMA, we can differentiate between DNA of intact cells (which PMA does not penetrate and is therefore amplifiable) versus relic DNA.

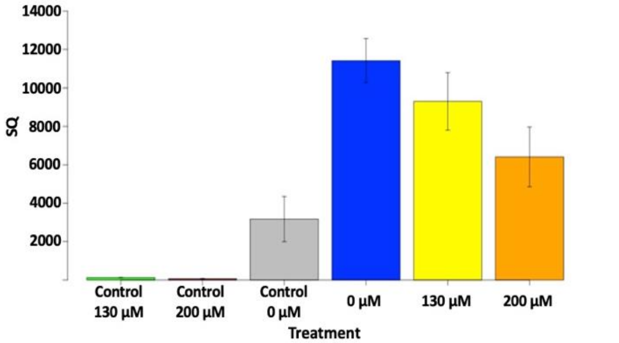

Figure 1. 16S rRNA gene abundance (SQ = starting quantity) for killed controls and viable soils with different propidium monoazide (PMA) concentrations, quantified via qPCR.

Before combining the PMA relic DNA protocol with SIP targeted genomics, we first tested the protocol for our soils—varying the PMA and soil concentrations, the photoactivation time and using a killed control (20 mins of boiling). Our initial results indicate that using different concentrations of PMA can result in different quantification results for the relic DNA present in viable soils (Fig 1). We have concluded that it will take significantly more effort to optimize the relic-DNA approach for our soils and may need to be customized for each distinctive soil.

Community RNA-seq: In the past decade, it has become increasingly clear that amplicon-based microbial ecology surveys that focus only on bacterial or fungal components of the soil microbiome miss a vast diversity of viruses and microfauna (including protists, nematodes and other soil invertebrates < 100 µm) who are significant contributors to ecological interactions and biogeochemical fluxes[12]. While high-throughput amplicon sequencing allows identification of multiple groups of soil organisms in parallel, PCR amplification has multiple biases, and the lack of a universal primer set means multiple primer sets are required to amplify taxonomically disparate groups. An alternative approach is to use amplification-independent methods for ribosomal community analysis, such as shotgun metagenomics or RNA sequencing (RNA-Seq). We used a tool called EMIRGE (developed in the Banfield Lab (UCB)[13, 14]) to reconstruct ribosomal sequences from a shotgun RNA sequence dataset of living and decomposing roots and to generate a marker gene-style abundance table for all present organisms, regardless of domain[15]. This "community RNA-Seq" analysis showed that when root litter was available, rhizosphere and bulk soil had significantly more Amoebozoa, which are potentially important yet often overlooked top-down drivers of

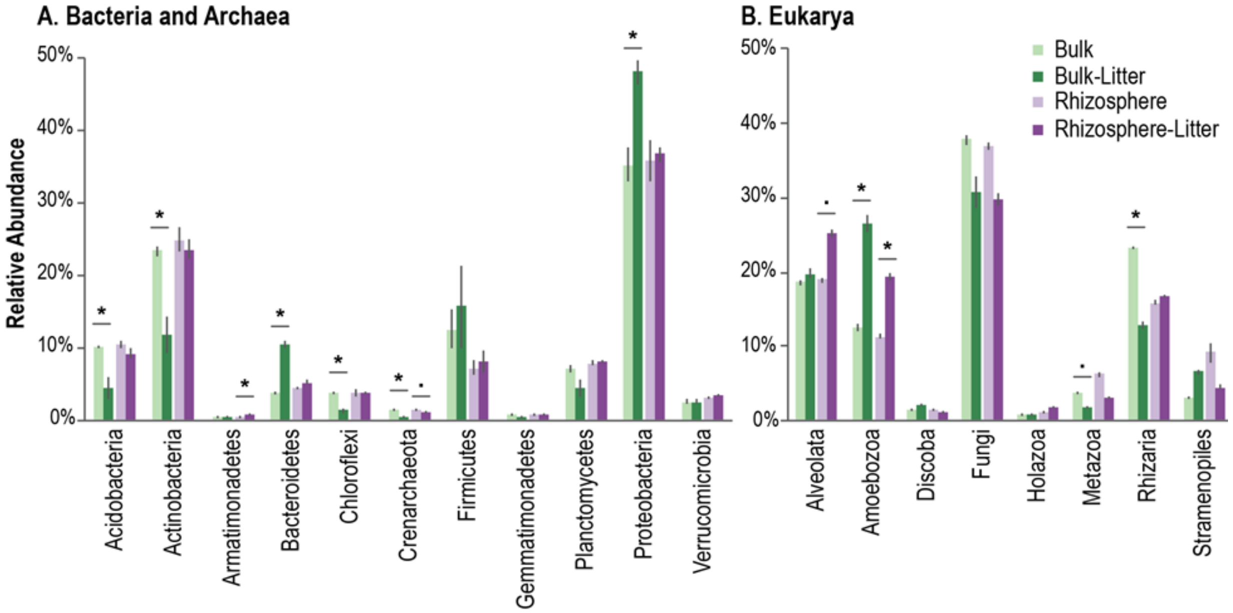

Figure 2: Community RNA-Seq analysis of litter decomposing microbes in the living rhizosphere verses bulk soil for (A) Bacteria and Archaea and (B) Eukarya[15].

detritusphere community dynamics and nutrient cycling (Fig 2). Bulk soil containing litter was depleted in Actinobacteria but had significantly more Bacteroidetes and Proteobacteria. Then, using the ***Chip-SIP*** isotope microarray technique developed at LLNL[16, 17], we found Actinobacteria preferentially incorporated litter relative to root exudates. Overall, our results emphasize that decomposition is a multi-trophic process involving cross-kingdom interactions. ***Community RNA-Seq*** is a particularly useful approach because it allows Bacteria, Archaea, and Eukarya to be studied simultaneously without amplification, and takes advantage of the naturally high coverage of ribosomal subunits used for taxonomic ID in RNA datasets (e.g. 16S, 18S, 28S), and yields greater sequencing depth of these regions than metagenomic sequencing.

Community networks and ecological modeling: Random Matrix Theory-based co-occurrence networks and ecological modeling of community assembly are also fruitful ways to explore composition and interactions in environmental microbiomes. While modules in these networks are not proof of an interaction, they can suggest shared niches and putative interactions. We used this approach to show that rhizosphere soil bacterial networks are far more complex than those in surrounding bulk soils, indicating a higher degree of interactions and niche-sharing. As plant roots grow, we observed increases in network complexity that were decoupled from community diversity[18]. In a related study, we found non-mycorrhizal fungi form increasingly complex networks with bacteria in rhizosphere soils, while arbuscular mycorrhizal fungi (AMF) form more network connections with bacteria in bulk soils[19]. In a third study, we used network and community assembly analysis to explore protist communities, again finding more complex networks in the rhizosphere compared to bulk soil[12]. These protists play varied ecological roles and their community assembly was primarily controlled by dispersal limitation and homogenous selection.

SIP-metagenomics for multi-domain analysis (bacteria, fungi, archaea, protists, viruses): All of life and many viruses encode their genomes on DNA. Using shotgun sequencing metagenomics, we can identify and, in many cases, reconstruct the genomes for these organisms in environmental samples. Many viruses, prokaryotes, and eukaryotes remain unculturable at this time, and other sequencing techniques rely on clade specific tags which may miss novel organisms and only provide phylogenic information. In contrast, metagenomics allows us to reconstruct near-complete genomes. When used in combination with SIP, we can ensure the genomes captured are of active (and thus more relevant) organisms. In SIP, a rare stable isotope (e.g. ^{18}O or ^{13}C) is incorporated into the genomes of growing organisms, making their DNA more dense, and allowing it to be separated in a gradient solution by centrifugation. Heavier DNA reflects more isotope uptake and thus more activity (Fig 3). SIP metagenomics not only improves metagenome quality but also provides valuable evidence for intertrophic interactions. In a recent study[20], we used SIP and genome-resolved metagenomics to demonstrate the community of bacteria which grew near roots (rhizosphere), those that could help the plant grow during stressful conditions, and plant pathogens. Then we identified micro-eukaryotes which were isotopically labelled and likely consumed the rhizosphere bacteria. We also reconstructed the complete genome of a virus which was parasitizing one of the pathogenic bacteria. This suggests that viral attack of soil bacteria could contribute to microbial death and be harnessed to control plant pathogens and to investigate soil carbon sequestration**.**

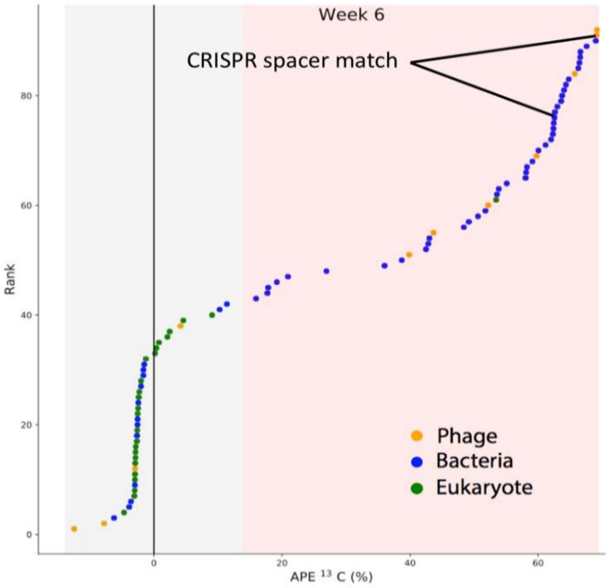

Figure 3. The rank of soil-derived phage genomes, bacterial genome bins, and scaffolds encoding eukaryotic 18S rRNA genes after six weeks of plant growth with $^{13}CO_2$, in order of their isotope enrichment or atom percent excess (APE). The gray region indicates unlabeled entities, and the red indicates labelled DNA. A labelled virus and host bacteria are indicated.

Benefits of metagenome (MAG)-focused analysis vs. amplicons: Because the majority of microbes cannot be cultured, culture-independent direct sequencing has become an essential tool for characterizing microbiomes. This is often conducted with sequence counts of amplified marker genes ('amplicons'), often the 16S/18S rRNA gene (bacteria, archaea, microfauna) or the ITS region (fungi). This approach requires far less sequencing per sample and less computational overhead. However, as microbial ecology moves beyond "who is there" to "what are they doing", the lack of functional information provided by amplicons (and the fact that amplicons miss some microbial diversity) has made shotgun metagenomics a more and more appealing way to gain a complete picture of the functional potential of a sample. By calculating the functions encoded on a contig and its depth of coverage, the metabolic capabilities of samples can be compared and further organized into pathways if the contigs are binned into MAGs.

Sample	*Total MAGs*	*With rps6*	*With 16S*
Tropical	326	232 (71.2%)	46 (14.1%)
Permafrost	153	93 (60.8%)	12 (7.8%)

Table 1. The number of rps6 and 16S rRNA genes matched with assembled MAGS in two soils.

Drawbacks of MAGs compared to amplicons include the increased computation and sequencing required. If a typical amplicon is 250 bp, and a typical MAG genome is 2.5 Mbp, then MAG sequencing would need ~10,000x more reads to reach the same sequencing depth. As this is not feasible, complex shotgun metagenome assemblies are typically under-sampled, yielding near-complete genomes for more abundant organisms, and partial or no genomes for less abundant organisms. Further, the greater sequencing effort of shotgun metagenomes necessitates we compromise and sequence only a subset of available samples. While using both technologies in parallel might appear to be a way to get both robust counts as well as functions, unfortunately, data integration problems make it difficult to match an amplicon with a MAG. The 16S rRNA gene typically used for amplicon sequencing does not bin well; it often occurs in multiple copies per genome, and current binning algorithms rely on coverage. We are exploring other widespread marker genes that are single-copy and bin well. For example, in two of our current datasets (tropical soils and permafrost[21-23]) the rps6 gene is more reliably found in our MAGs than a 16S rRNA gene (Table 1).

New tools for metagenome-based genome curation: For our SFA, the recovery of accurate genomes (MAGs) from genome-resolved metagenomics is absolutely essential. Currently, a single sequencing read is ~8000x smaller than the entire genome of a typical microbe. To reconstruct the thousands (potentially millions) of distinct microbial genomes in a sample, these sequencing reads must be assembled into larger contiguous sequences. Despite major advances, assemblers still produce significant errors. These can be corrected, but the process currently requires manual human-guided curation that necessitates a huge amount of time and expertise. Typical metagenomic projects from complex environments produce thousands of genomes, and human-guided curation is not possible on such a massive scale. To overcome the bottlenecks of human-guided genome curation, our SFA is developing a computational suite to automatically identify and repair metagenomic assembly errors (Fig 4). The proof-of-concept is now being packaged into a software solution 'FixAME' for KBase and can effectively repair errors in thousands of sequences in hours.

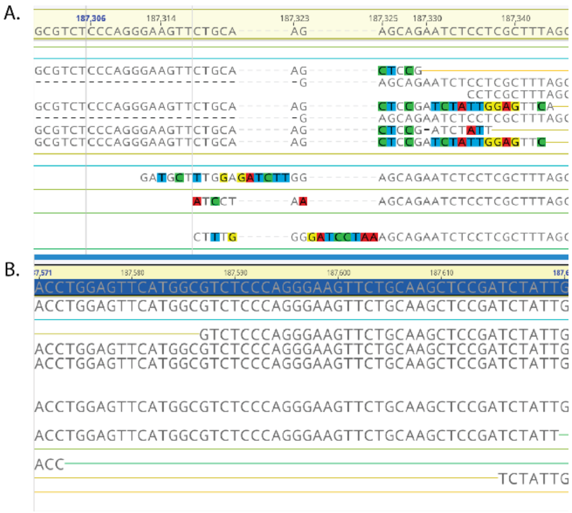
A.
B.
GCGTCTCCCAGGGAAGTTCTGCA AG AGCAGAATCTCCTCGCTTTAGC
ACCTGGAGTTCATGGCGTCTCCCAGGGAAGTTCTGCAAGCTCCGATCTATTG

Figure 4. A) Read alignment of a phage assembled sequence[38]. Aligned reads show many disagreements with the assembled sequenced, represented by mismatches and gaps. B) After applying FixAME to the same sequence. Sequencing reads align to the region with proper paired-end support and no mismatches or gaps—indicating the region is repaired and error-free.

Don't forget the small microbes: Another microbial group that is frequently missed by traditional microbiome surveys are extremely small cells and taxa with few 16S copies. Using an approach designed for recovery of viral particles, we targeted small and overlooked microbes, including the Candidate Phyla Radiation (CPR) bacteria and DPANN archaea (an acronym of the names of the first included archaea phyla). We size-fractionated and concentrated small particles (< 0.2 μm) from soil to sample genomes that were absent from non-size fractionated metagenomes (Fig 5). We achieved CPR and DPANN enrichments of 100- to 1000-fold compared to bulk soil. We estimate that there are approximately 1 to 100 cells from each of these lineages per gram of soil, suggesting that this approach provides a window onto the soil rare biosphere. The organisms we detected and created MAGs for include Doudnabacteria (SM2F11) and Pacearchaeota genomes, organisms rarely reported in soil, as well as Saccharibacteria, Parcubacteria and Microgenomates[24].

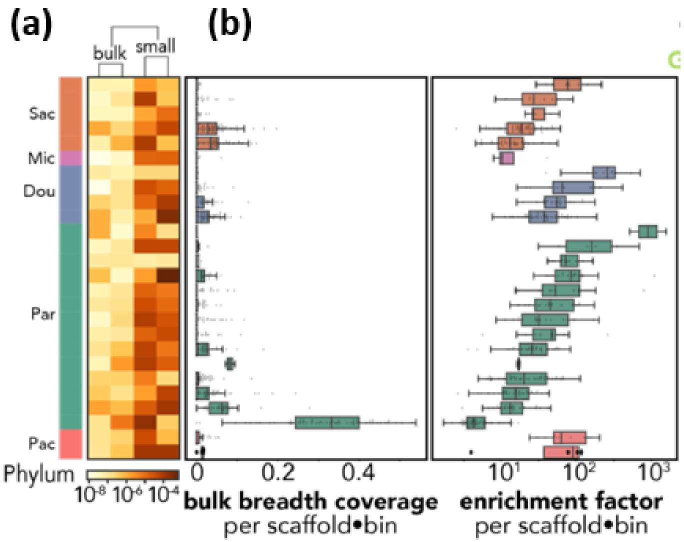

Figure 5. Enrichment and metabolic profiles of CPR in soil concentrate metagenomes. (a) Comparison of metagenomic reads after concentrating small particles from soil vs. bulk soil metagenomes. (b) Heatmap showing relative abundance of 26 organisms by phylum (Sac: Saccharibacteria, Mic: Microgenomates, Dou: Doudnabacteria, Par: Parcubacteria, Pac: Pacearchaeota) across bulk metagenomes and concentrate metagenomes.

II. Viral Communities

Viruses are ubiquitous and the most abundant biological entities on Earth, infecting all living organisms. However, our current understanding of viruses in terrestrial habitats is limited by the intrinsic complexity of the soil matrix, the immense diversity of microbes, and viral recalcitrance to laboratory cultivation, which often prevents the identification of specific viruses. In the last decade, virus ecology has evolved from gene-based to genome-based, and has begun to identify the broad diversity of virus communities in many ecosystems. Viruses are thought to impact carbon cycling by controlling microbial communities via predation, transferring genes from one host to another, and metabolically reprogramming their host cells via regulatory take-over and directly-encoded auxiliary metabolic genes (AMGs). While viruses are abundant (as high as 10^9 particles per gram of soil) in soils, relatively little is known about virus community composition in soils.

There are two major ways in which virus communities from soil can be studied: either from bulk metagenomes and transcriptomes or from virus-enriched 'virome' datasets, where filtration steps and DNase treatments reduce host nucleic acids, leaving behind the virus fraction (Fig 6). We are actively investigating the differences between these two approaches. Upcoming challenges in this field include a) improved recovery of virus populations, b) determining whether recovered viruses are actively replicating, and c) how to accurately assign host linkage(s) to each virus genome within mixed community metagenomes.

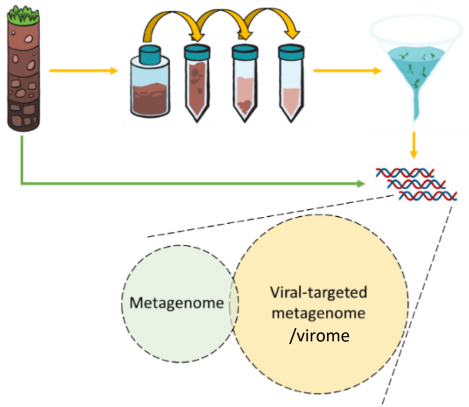

Figure 6. Schematic comparing metagenomes and viral-targeted metagenomes (virome). A metagenome describes all the DNA in a sample (green). A virome requires pre-processing and filtering (typically through a 0.22 μm filter) and typically provides increased viral sequence information (yellow).

Comparing viral detection in metagenomes and viromes: To better understand the difference between viruses from bulk metagenomes and virus-enriched metagenomes (viromes), we helped to compare 87 metagenomes and five viral-targeted metagenomes sequenced from a boreal peatland[25]. Each approach yielded unique viruses, but on a per-sample basis, viral operational taxonomic unit (vOTU) recovery was 32 times higher from the viromes compared to the metagenomes. A comparison of datasets from the two approaches suggested the metagenomes were well-sampled but the viromes were undersampled, suggesting that the viromes could have yielded even more viral sequences with deeper sequencing, providing access to the rare virosphere.

Virome Diversity: We have also compared viral communities across multiple ecosystems, including CA grasslands with a range of climates and subtropical soils from Puerto Rico, to assess viral diversity and interactions with microbial communities in diverse soils[26]. From these samples, we identified 36,867 unique viruses (vOTUs; viral contigs dereplicated at 95% average nucleotide identity and 80% coverage). To our knowledge, this cross-site study is currently the largest dataset of viral sequences from individual soil virome samples. We continue to find that soil viruses are vastly undersampled. For example, our newly generated database is almost as large as the current database of cultivated and uncultivated viruses (IMG/VR v3) which contains only 43,586 vOTUs from 1994 genomes or metagenomes from soils.

We compared our viruses to terrestrial viruses from a global virome study, permafrost viruses, and viral genomes from the online database RefSeq by generating a network of gene-sharing clusters that approximate 'genera'[27]. Our soil viruses contained 2,745 genera, of which ~50% were novel (Fig. 7A). In our cross-site virome study, the vast majority of viruses were exclusive to each site (84%-99%) (Fig. 7B). Viral communities were strongly separated by location (Fig. 7C), with soil moisture and temperature as the strongest drivers of viral community structure. To evaluate the potential impacts of these viruses, we characterized viral-encoded auxiliary metabolic genes (AMGs) using the informatics tool, DRAM-v[28]. We identified 249 distinct AMGs from diverse metabolic pathways. Many of these genes had distinct ecologies, which may reflect variations in carbon metabolic limitations to virus infection.

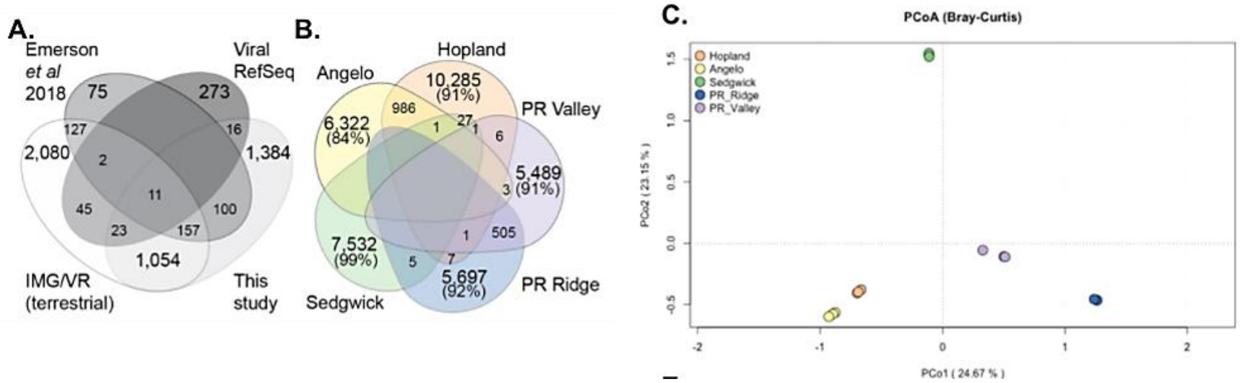

Figure 7. A. Distribution of viral genera across publicly available soil/terrestrial datasets. B. Number and overlap of vOTUs across our 5 soil sites in CA and PR. C. Principal coordinates analysis (PCoA) ordination of viral communities from each sample.

Identifying RNA viruses and their potential hosts with metagenomics and metatranscriptomics: The vast majority of environmental virus surveys focus on DNA viruses; by comparison, RNA viruses are an understudied and unknown player in environmental samples. In a recent publication[29], we investigated the diversity and ecology of RNA viruses and their hosts. To conduct this analysis, we collaborated with the Join Genome Institute (JGI) to sequence mRNA from soil samples and then used bioinformatic algorithms to reconstruct RNA viral genomes and map to a site-specific metagenome[30]. This technique had never been applied to soil and is infrequently used in environmental samples. We identified a large diversity of soil RNA viruses, indicating that soils may be reservoirs for novel RNA viruses (Fig 8). To understand which organisms served as viral hosts, we reconstructed key genes (from the metatranscriptome) for the bacteria,

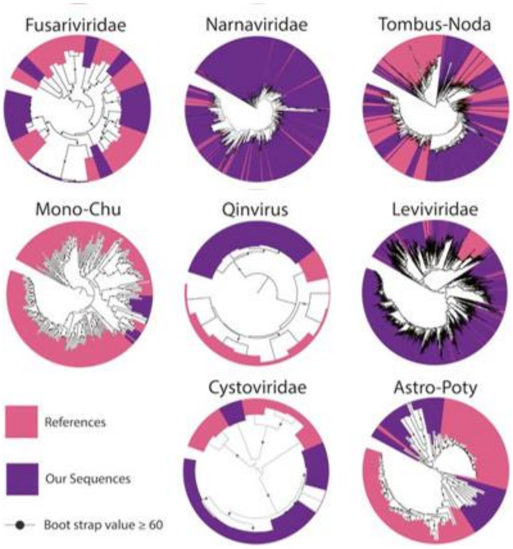

Figure 8. Phylogenetic trees representing clades of RNA viruses identified in our California annual grassland experimental soil. Within each tree, the RdRp sequences we identified are colored purple and previously described sequences are in pink.

archaea, fungi, protists, and micro-eukaryotes present. Based on the phylogeny of the identified RNA viruses, it appears that many of the fungi we found were viral hosts. Improving techniques and tools for environmental monitoring of RNA viruses is especially important right now, given that coronaviruses (the cause of SARS, MERS and COVID-19) are RNA viruses.

Tracking viruses under different soil conditions with SIP-metagenomics: Isotopic tracers have been recognized for over a century as a powerful approach for biological systems; in the field of viral biology, isotope labeling has facilitated many important discoveries. Viruses preferentially use extracellular nutrients to build their progeny. This means that isotopically labeled substrates can be added to an environmental sample, and active viruses will become isotopically enriched (‘labeled’). We have used this approach, stable isotope probing (SIP), in combination with metagenomic sequencing to track and characterize active microbes and viruses in several soil systems.

In the first study, we used ‘heavy water’ ($H_2^{18}O$) SIP-metagenomics to study active viruses in permafrost-associated peatland soils incubated under winter-like conditions (anoxic and subzero)[21]. We assembled ~52,000,000 contigs from 23 SIP-metagenomes and identified 153 microbial populations (MAGs)[23] and 332 viral populations (vOTUs). Most of the vOTUs came from the isotope-enriched DNA fractions, indicating they were active during the soil incubation. Using host matching techniques involving CRISPR sequences and whole-genome similarity (Fig 9), we linked 33% of the active vOTUs to 51% of the active MAGs. Over the year-long incubation, the active virus community richness and abundance dramatically changed. This is the first application of ^{18}O SIP to label viruses. Our data show a diverse array of active microbes and viruses in anoxic subfreezing soil, revealing an ongoing arms race over winter months, where viruses play an important role in shaping microbial populations and limiting microbial metabolic outputs.

In a second study, we tracked viruses infecting microbes that degrade organic matter in a wet tropical forest soil that naturally experiences dynamic redox conditions. We incubated soils with ^{13}C-plant biomass under 4 redox treatments. From over 85 SIP-fractionated metagenomes, we identified 326 MAGs and 640 vOTUs. SIP-fraction samples recovered 7% more vOTUs, these would have been missed in a traditional ‘bulk’ metagenome. A comparison of the redox treatment effects indicated that viral diversity was highest in the oxic samples and decreased in soils with lower O_2 exposure. In these soils, only 27% of the vOTUs were active overall, and 16% were only active in the anoxic samples. Almost 30% of the vOTUs were able to be linked to the 326 MAGs we assembled from SIP metagenomes.

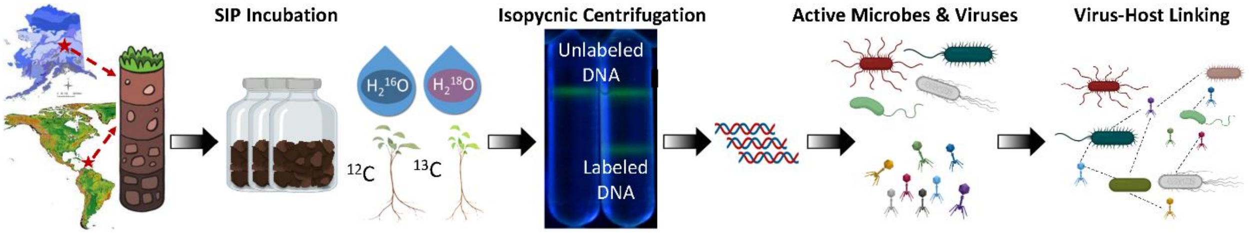

Figure 9. Generalized workflow for a SIP-metagenomics study, to identify and track active viruses, and then link them to microbial hosts using CRISPR sequence matching and whole-genome similarity.

In a third SIP-metagenome study, we analyzed active viruses from our SFA's three focal sites in California grasslands that occur along a rainfall gradient ('Hopland', 'Angelo', and 'Sedgwick')[31]. Soils were incubated with 'heavy water' ($H_2{}^{18}O$), then we determined the composition, functional potential and activity of soil virus communities using a new custom virus-host linkage workflow that we believe will improve the number of host predictions compared to current strategies. This allowed us to capture sufficient isotope-labelled viral DNA to calculate per-genome activity metrics (expressed as 'atom fraction excess' or AFE) for >8,000 vOTUs. We found the fraction of active viruses can vary significantly, ranging from 25% to 75% (Fig 10A). From this same dataset, >400 high-quality MAG bins were recovered; these were used as a training dataset to infer virus-host associations using multiple compositional features between bacterial and virus genomes. Here, we predicted high-confidence host linkages (up to family level) for ~50% of the vOTUs. Actinobacteria -specifically Mycobacteria – were the dominant host taxa across all three soils (Fig 10B). These predictions, combined with virus activity profiles, indicate that viruses are actively preying upon key microbial taxa known to have fundamental roles in soil ecosystem function.

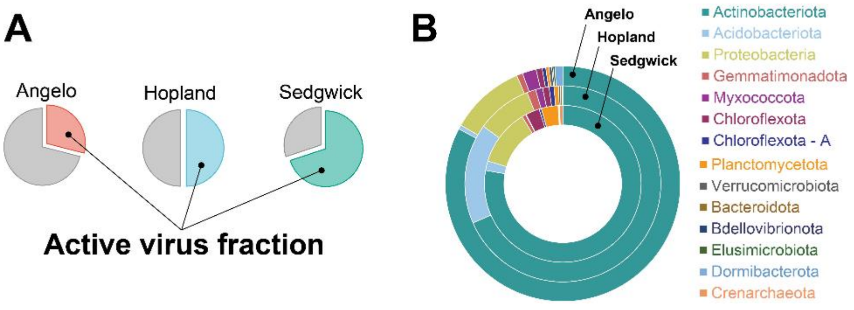

Figure 10. Virus activity profiles across three CA annual grassland soil sites. (A) Per-soil fraction of active viruses within the total number of identified viruses. (B) Doughnut chart depicting the proportion of host phyla per site.

Tool development for viral community ecology (VirION2, iVirus, VirMatcher): Though poorly understood in soils (e.g. compared to oceans), soil viruses are abundant and very likely have a large impact on carbon cycling microbes and their metabolisms[32]. However, studying viruses requires a very different informatics toolkit than is available for studying microbes. We have focused on making advances in this space in three ways: 1) improving data generation capabilities, 2) democratizing the existing 'iVirus' analytical toolkit by implementing its core components on DOE's KnowledgeBase (KBase), and 3) establishing a new analytic that enables better host prediction for newly discovered viruses.

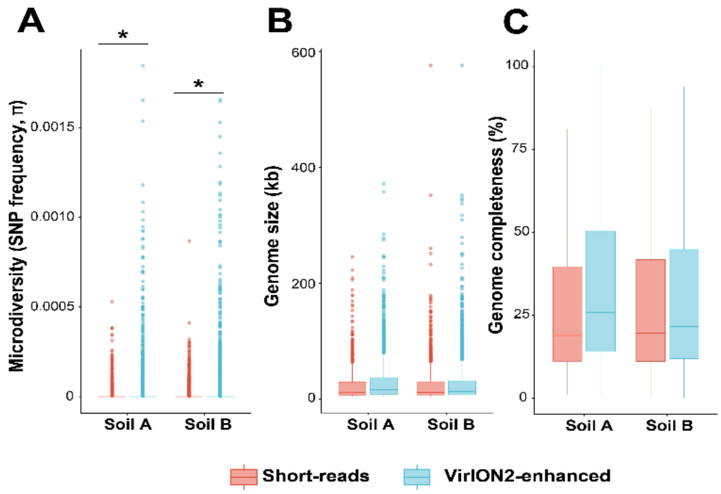

Figure 11. Genomic trait comparisons between short-read versus VirION2-enhanced (i.e., containing long-read) viromes. (A) microdiversity (B) genome size and (C) genome completeness.

Our first effort was to develop better underlying sequence data. Specifically, we made new improvements to a long-read sequencing protocol – VirION2[33] – making significant advances over our first version, VirION[34]. Now, the DNA input requirements are only 1ng and median read length is now ~7,000bp (a 100-fold reduction and a 76% increase from the original protocol, respectively). Up to 22% of the most abundant viruses in the samples could be recovered solely with long-reads. Further, by combining corrected long-reads along with high-accuracy (99.97%) short-reads from the same virome sample, we were able to recover up to 30% of the viruses in the community. As a proof of concept, we generated long-read data from several soil viromes, and found that virus microdiversity (i.e., intra-population, per genome SNP frequency) was significantly greater compared to short-read only data (Fig 11A). In addition, a greater number of longer (Fig 11B) and more complete (Fig 11C) virus genomes were recovered from the samples.

In a second project, we built out an analytical toolkit for researchers using the DOE Knowledgebase (KBase). These efforts leverage years of past effort to develop 'iVirus', an ecosystem of software apps, datasets, and resources on the CyVerse Cyberinfrastructure[35]. Here, we ported several critically missing apps to KBase including a virus identification tool (VirSorter[36]) and a virus classification tool (vConTACT2[27]) and provided a narrative and webinar to train researchers in the existing viral ecogenomic workflow available at KBase (Fig 12). These efforts, and recent upgrades to iVirus at CyVerse are described in a nearly-complete manuscript[37]. Future improvements will include bringing in other virus identification tools (MARVEL, DeepVirFinder, VIBRANT, VirSorter2), database updates as new data types become available, linkages through narratives to diverse protocols.io based informatics documentation, and interactive tables that incorporate Krona and other graphics to better leverage KBase's visualization capabilities.

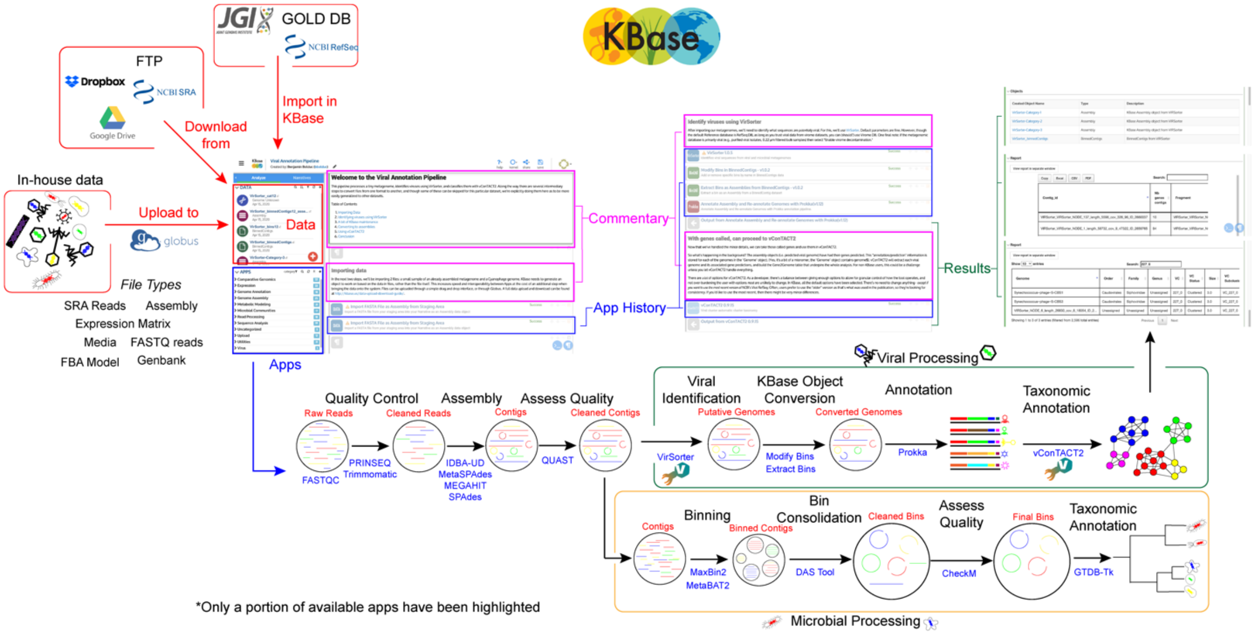

Figure 12. Overview of the viral ecogenomics workflow in the KBase ecosystem. Data sources are outlined in red, with the data window from the KBase narrative boxed in red, available apps boxed in blue, commentary boxed in purple, and results highlighted in green. Below is a pipeline starting from raw reads (data is in red font), tools used (in blue font) and processed to assembly. Upon assembly, viral identification splits to viral or microbial processing. Each stage in the pipeline is a distinct cell in the narrative, providing historical information associated with the app's execution, and results from the analysis.

A third focus has been to improve the *in silico* ability to predict hosts for the thousands to hundreds of thousands of new viruses discovered in the average study. The new host prediction tool aggregates existing *in silico* capabilities within a probabilistic scoring framework to provide not only a host prediction, but also a systematically evaluated "confidence score" for the result. This new tool – VirMatcher – is complete and tested and currently being incorporated into KBase as a new capability.

Summary

The LLNL *Microbes Persist* Soil Microbiome SFA uses a multi-domain approach to identify the microbial, microfauna and viral inhabitants of soil ecosystems, designed to provide a comprehensive understanding of biotic interactions, ecophysiological traits, and the fate of microbiome biomass organic carbon. In both our empirical research and methods development, we are moving beyond traditional assessments of microbial communities by pairing stable isotope probing and assessments of inactive (relic) nucleic acids with metagenomic and metatranscriptomic surveys. This allows us to differentiate between actively growing microorganisms, dead and degraded DNA, and DNA/RNA from all taxonomic groups (including viruses) that make up the soil microbiome. Our approaches capture the diversity of both commonly described microbial communities (i.e., bacteria, archaea, fungi), but also micro-eukaryotes, ultra-small prokaryotes and indigenous viruses that reside in distinct soil physical habitats. This community composition information is foundational to our efforts to understand microbial traits, ecological interactions, and genomic potential in soil microbiomes. Stable isotope probing enabled approaches are particularly key to our efforts, giving us an unprecedented picture of the most relevant taxa in soil ecosystems. Of equal importance are new informatics applications we are developing, including a computational suite to automatically identify recovered genomes, detect key functional genes, link intertrophic interactions, and predict ecological drivers on community structure. These new tools not only help us to develop a microbiome-informed predictive understanding of soil carbon persistence but also provide valuable resources to the broader scientific community.

References

1. Blazewicz, S. J.; Hungate, B. A.; Koch, B. J.; Nuccio, E. E.; Morrissey, E.; Brodie, E. L.; Schwartz, E.; Pett-Ridge, J.; Firestone, M. K., Taxon-specific microbial growth and mortality patterns reveal distinct temporal population responses to rewetting in a California grassland soil. *The ISME Journal* **2020,** doi.org/10.1038/s41396-020-0617-3, 1-13.
2. Shi, S. J.; Nuccio, E.; Herman, D. J.; Rijkers, R.; Estera, K.; Li, J. B.; da Rocha, U. N.; He, Z. L.; Pett-Ridge, J.; Brodie, E. L.; Zhou, J. Z.; Firestone, M., Successional Trajectories of Rhizosphere Bacterial Communities over Consecutive Seasons. *Mbio* **2015,** *6*, (4). DOI:10.1128/mBio.00746-15
3. Shi, S.; Herman, D. J.; He, Z.; Pett-Ridge, J.; Wu, L.; Zhou, J.; Firestone, M. K., Plant roots alter microbial functional genes supporting root litter decomposition. *Soil Biology and Biochemistry* **2018,** *127*, 90-99. doi.org/10.1016/j.soilbio.2018.09.013
4. Pett-Ridge, J.; Firestone, M. K., Using stable isotopes to explore root-microbe-mineral interactions in soil. *Rhizosphere* **2017,** *3*, 244-253. https://doi.org/10.1016/j.rhisph.2017.04.016
5. Finley, B. K.; Hayer, M.; Mau, R. L.; Purcell, A. M.; Koch, B. J.; van Gestel, N. C.; Schwartz, E.; Hungate, B. A., Microbial taxon-specific isotope incorporation with DNA quantitative stable isotope probing. In *Stable Isotope Probing*, Springer: 2019; pp 137-149. https://doi.org/10.1007/978-1-4939-9721-3_11
6. Sieradzki, E. T.; Koch, B. J.; Greenlon, A.; Sachdeva, R.; Malmstrom, R. R.; Mau, R. L.; Blazewicz, S. J.; Firestone, M. K.; Hofmockel, K.; Schwartz, E.; Hungate, B. A.; Pett-Ridge, J., Measurement error and resolution in quantitative stable isotope probing: implications for experimental design, **2020,** *5*, e00151-20. https://doi.org/10.1128/mSystems.00151-20.
7. Koch, B. J.; McHugh, T. A.; Hayer, M.; Schwartz, E.; Blazewicz, S. J.; Dijkstra, P.; van Gestel, N.; Marks, J. C.; Mau, R. L.; Morrissey, E. M.; Pett-Ridge, J.; Hungate, B. A., Estimating taxon-specific population dynamics in diverse microbial communities. *Ecosphere* **2018,** *9*, (1), e02090-15. https://doi.org/10.1002/ecs2.2090
8. Li, J.; Mau, R. L.; Dijkstra, P.; Koch, B. J.; Schwartz, E.; Liu, X.-J. A.; Morrissey, E. M.; Blazewicz, S. J.; Pett-Ridge, J.; Stone, B. W.; Hayer, M.; Hungate, B. A., Predictive genomic traits for bacterial growth in culture versus actual growth in soil. *The ISME Journal* **2019**. https://doi.org/10.1038/s41396-019-0422-z
9. Hungate, B. A.; Mau, R. L.; Schwartz, E.; Caporaso, J. G.; Dijkstra, P.; van Gestel, N.; Koch, B. J.; Liu, C. M.; McHugh, T. A.; Marks, J. C.; Morrissey, E. M.; Price, L. B., Quantitative microbial ecology through stable isotope probing. *Applied and Environmental Microbiology* **2015,** *81*, (21), 7570-7581. https://doi.org/10.1128/AEM.02280-15
10. Carini, P.; Marsden, P. J.; Leff, J. W.; Morgan, E. E.; Strickland, M. S.; Fierer, N., Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nature microbiology* **2016,** *2*, 16242.
11. 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* **2014,** *95*, (5), 1162-1172.
12. Ceja-Navarro, J. A.; Wang, Y.; Arellano, A.; Ramanculova, L.; Yuan, M.; Byer, A.; Craven, K.; Saha, M.; Brodie, E.; Pett-Ridge, J.; Firestone, M. K., Protist diversity and network complexity in the rhizosphere are dynamic and changing as the plant develops. *Microbiome* **in review**.
13. Miller, C. S.; Baker, B. J.; Thomas, B. C.; Singer, S. W.; Banfield, J. F., EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome biology* **2011,** *12*, (5), R44. https://doi.org/10.1186/gb-2011-12-5-r44
14. Miller, C. S.; Handley, K. M.; Wrighton, K. C.; Frischkorn, K. R.; Thomas, B. C.; Banfield, J. F., Short-read assembly of full-length 16S amplicons reveals bacterial diversity in subsurface sediments. *PLoS ONE* **2013,** *8*, (2), e56018.
15. Nuccio, E. E.; Nguyen, N. H.; Rocha, U. N. d.; Mayali, X.; Bougoure, J.; Weber, P.; Brodie, E.; Firestone, M.; Pett-Ridge, J., Community RNA-Seq: Multi-kingdom responses to living versus decaying root inputs in soil. *bioRxiv (ISME Communications,* in review*)* **2021,** https://doi.org/10.1101/2021.01.12.426429.

16. Mayali, X.; Weber, P. K.; Brodie, E. L.; Mabery, S.; Hoeprich, P. D.; Pett-Ridge, J., High-throughput isotopic analysis of RNA microarrays to quantify microbial resource use. *ISME J* **2012,** *6*, (6), 1210-1221. https://doi.org/10.1038/ismej.2011.175
17. Mayali, X.; Weber, P. K.; Pett-Ridge, J., Taxon-specific C:N relative use efficiency for amino acids in an estuarine community. *FEMS Microbiology Ecology* **2013,** *83*, (2), 402-412. https://doi.org/10.1111/j.1574-6941.12000.x
18. Shi, S.; Nuccio, E. E.; He, Z.; Zhou, J.; Firestone, M. K., The interconnected rhizosphere: High network complexity dominates rhizosphere assemblages. *Ecology Letters* **2016,** *19*, (8), 926-936 *equal contribution. https://doi.org/10.1111/ele.12630
19. Yuan, M. M.; Kakouridis, A.; Starr, E.; Nguyen, N.; Shi, S.; Pett-Ridge, J.; Nuccio, E.; Zhou, J.; Firestone, M., Fungal-bacterial co-occurrence patterns differ between AMF and non-mycorrhizal fungi across space and time. *mBio* **in review**.
20. Starr, E.; Shi, S.; Blazewicz, S.; Probst, A.; Herman, D.; Firestone, M.; Banfield, J., Stable isotope informed genome-resolved metagenomics uncovers potential trophic interactions in rhizosphere soil. *bioRxiv (Microbiome, in review)* **2020** doi.org/10.1101/2020.08.21.262063.
21. Trubl, G.; Kimbrel, J. A.; Liquet-Gonzalez, J.; Nuccio, E. E.; Weber, P. K.; Pett-Ridge, J.; Jansson, J. K.; Waldrop, M. P.; Blazewicz, S. J., Tracking active virus population dynamics in Arctic peat soil via H218O stable isotope probing metagenomics. *ISME Journal* **in prep.**
22. Campbell, A.; Bhattacharyya, A.; Tfaily, M.; Thompson, A.; Chu, R.; Pasa-Tolic, L.; Lin, Y.; Silver, W.; Nico, P.; Pett-Ridge, J., Impacts of dynamic soil redox on tropical soil microbiomes and biogeochemical transformations. *ISME Journal* **in prep.**
23. Blazewicz, S. J.; White, R. A.; Kimbrel, J. A.; Tas, N.; Euskirchen, E. S.; McFarland, J.; Jansson, J. K.; Waldrop, M. P., Life in ice: microbial fermentation controls over-winter carbon release from a collapsed permafrost bog. *ISME Journal* **in prep.**
24. Nicolas, A. M.; Jaffe, A. L.; Nuccio, E. E.; Taga, M. E.; Firestone, M. K.; Banfield, J. F., Unexpected diversity of CPR bacteria and nanoarchaea in the rare biosphere of rhizosphere-associated grassland soil. *bioRxiv* **2020**. https://doi.org/10.1101/2020.07.13.194282
25. ter Horst, A. M.; Santos-Medellín, C.; Sorensen, J. W.; Zinke, L. A.; Wilson, R. M.; Johnston, E. R.; Trubl, G. G.; Pett-Ridge, J.; Blazewicz, S. J.; Hanson, P. J.; Chanton, J.; Schadt, C.; Kostka, J.; Emerson, J., Minnesota peat viromes reveal terrestrial and aquatic niche partitioning for local and global viral populations. *bioRxiv (Microbiome, in review)* **2020**. https://doi.org/10.1101/2020.12.15.422944
26. Sun, C. L.; Zablocki, O.; Greenlon, A.; Zayed, A. A.; Nicolas, A.; Nuccio, E.; Solden, L.; Bolduc, B.; Firestone, M.; Banfield, J. F.; Blazewicz, S.; Pett-Ridge, J.; Sullivan, M. B., Novel soil viruses are soil specific and contain genes involved in carbon cycling. **in prep.**
27. Jang, H. B.; Bolduc, B.; Zablocki, O.; Kuhn, J. H.; Roux, S.; Adriaenssens, E. M.; Brister, J. R.; Kropinski, A. M.; Krupovic, M.; Lavigne, R.,… Sullivan, M.B. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nature biotechnology* **2019,** *37*, (6), 632-639. https://doi.org/10.1038/s41587-019-0100-8
28. Shaffer, M.; Borton, M. A.; McGivern, B. B.; Zayed, A. A.; La Rosa, S. L.; Solden, L. M.; Liu, P.; Narrowe, A. B.; Rodríguez-Ramos, J.; Bolduc, B.; Gazitúa, M. C.; Daly, R. A.; Smith, G. J.; Vik, D. R.; Pope, P. B.; Sullivan, M. B.; Roux, S.; Wrighton, K. C., DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic acids research* **2020,** *48*, (16), 8883-8900. https://doi.org/10.1093/nar/gkaa621
29. Starr, E. P.; Nuccio, E. E.; Pett-Ridge, J.; Banfield, J. F.; Firestone, M. K., Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil. *Proceedings of the National Academy of Sciences* **2019,** *116*, (51), 25900-25908. https://doi.org/10.1073/pnas.1908291116
30. Nuccio, E. E.; Starr, E.; Karaoz, U.; Brodie, E. L.; Zhou, J.; Tringe, S. G.; Malmstrom, R. R.; Woyke, T.; Banfield, J. F.; Firestone, M. K., Pett-Ridge, J. Niche differentiation is spatially and temporally regulated in the rhizosphere. *The ISME Journal* **2020**, 1-16. https://doi.org/10.1038/s41396-019-0582-

x

31. Zablocki, O.; Greenlon, A.; Zayed, A.; Blazewicz, S.; Pett-Ridge, J.; Sullivan, M., Phage activity in soils revealed by single isotope probing. **in prep.**
32. Emerson, J. B.; Roux, S.; Brum, J. R.; Bolduc, B.; Woodcroft, B. J.; Jang, H. B.; Singleton, C. M.; Solden, L. M.; Naas, A. E.; Boyd, J. A.,…Sullivan, M.B. Host-linked soil viral ecology along a permafrost thaw gradient. *Nature microbiology* **2018,** *3*, (8), 870-880. https://doi.org/10.1038/s41564-018-0190-y
33. Zablocki, O.; Michelsen, M.; Burris, M.; Solonenko, N.; Warwick-Dugdale, J.; Ghosh, R.; Pett-Ridge, J.; Sullivan, M. B.; Temperton, B., VirION2: a short-and long-read sequencing and informatics workflow to study the genomic diversity of viruses in nature. *bioRxiv* **2020**. https://doi.org/10.1101/2020.10.28.359364
34. Warwick-Dugdale, J.; Solonenko, N.; Moore, K.; Chittick, L.; Gregory, A. C.; Allen, M. J.; Sullivan, M. B.; Temperton, B., Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. *PeerJ* **2019,** *7*, e6800. https://doi.org/10.7717/peerj.6800
35. Bolduc, B.; Youens-Clark, K.; Roux, S.; Hurwitz, B. L.; Sullivan, M. B., iVirus: facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure. *The ISME journal* **2017,** *11*, (1), 7-14. DOI: 10.1038/ismej.2016.89
36. Roux, S.; Enault, F.; Hurwitz, B. L.; Sullivan, M. B., VirSorter: mining viral signal from microbial genomic data. *PeerJ* **2015,** *3*, e985. https://doi.org/10.7717/peerj.985
37. Bolduc, B.; Guo, J.; Zablocki, O.; Dehal, P.; Wood-Charlson, E.; Pett-Ridge, J.; Vaughn, M.; Merchant, N.; Arkin, A.; Sullivan, M. B., iVirus 2.0: an expanding suite of viral ecology tools and data. **in prep.**
38. Al-Shayeb, B.; Sachdeva, R.; Chen, L.-X.; Ward, F.; Munk, P.; Devoto, A.; Castelle, C. J.; Olm, M. R.; Bouma-Gregson, K.; Amano, Y.,… Banfield, J.F. Clades of huge phages from across Earth's ecosystems. *Nature* **2020,** *578*, (7795), 425-431. https://doi.org/10.1038/s41586-020-2007-4

This work was conducted at Lawrence Livermore National Laboratory, supported by DOE OBER Genomic Sciences award SCW1632 and conducted under the auspices of DOE Contract DE-AC52-07NA27344.

IM# LLNL-TR-818366