BSSD 2021 Performance Metric Q3

Goal: Develop and apply computational approaches to understand microbiome function in environmental samples

Q3 Target: Describe progress on computational approaches used to analyze microbial activities in environmental microbiomes

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 formation of soil carbon (C). Our SFA research program is now four years old; it evolved from previously-funded BSSD projects in the Firestone (UCB), Banfield (UCB), Sullivan (OSU) and Hungate (NAU) labs. We use stable isotope probing (SIP) in combination with 'omics analyses to measure how dynamic water regimes shape activity of individual microbial populations *in situ* and how their ecophysiological traits 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 how soil microbiomes shape the fate of soil C. Here we focus on computational approaches (and applications) that advance our understanding of dynamics in complex soil microbiomes.

Optimizing Quantitative Stable Isotope Probing

Stable isotope probing (SIP) is one of the few approaches that can identify the ecophysiology of active microorganisms in their native environments, making it one of the most powerful techniques in microbial ecology. Broadly speaking, SIP refers to any technique where microorganisms that have actively consumed substrates enriched in rare stable isotopes (e.g. C, N, O) are identified based on the resulting isotopic enrichment of their nucleic acids, proteins, and metabolites. Density gradient SIP is the culture-independent gold standard for directly linking sequence to function in complex microbial communities[1]. When a microbe consumes a substrate enriched with a heavy isotope, the cellular components of that cell also become labeled in the heavy isotope. Density gradient SIP takes advantage of the increased density of microbial nucleic acids (due to assimilation of heavy isotopes), using a density gradient to separate the heavy (labeled) nucleic acids from lighter (unlabeled) ones. Isolated heavy nucleic acids can then be characterized to identify the organisms that actively assimilated substrates of interest.

Our SFA team has pioneered new SIP computational approaches that quantify element fluxes with high taxonomic resolution. In particular, quantitative stable isotope probing (qSIP) developed at NAU with LLNL help, is the isopycnic separation of nucleic acids in cesium chloride combined with a mathematical model to quantify isotope enrichment[2-5]. With qSIP we measure growth rates of individual taxa or viruses in complex soil communities using O-labeled water as a universal substrate that is used by all actively growing organisms. As described below, our group continues to make efforts to improve quantitative accuracy of qSIP calculations.

qSIP precision and statistical power

One of the standard perceptions of many SIP practitioners is that increased resolution (i.e., more density fractions) leads to improved detection of active organisms represented by amplicon variants. However, as the field transitions from 16S-rRNA amplicon studies to sequencing metagenomes from density fractions, it can be both financially and computationally prohibitive to

run a SIP study with both robust replication, and many density fractions. We simulated results from multiple experimental datasets to represent the effects of different fraction resolutions, and found diminishing returns when more than nine fractions are used. We also showed that reduction of the number of fractions has little impact on sensitivity and specificity as long as a detection limit is kept at a minimum of 0.005 g*ml^{-1} (equivalent to 9 atom % enrichment of ^{13}C). Another SIP paradigm is that most of the variability is generated between batches (tubes spun at different times in the ultracentrifuge). However, we showed that the variability between batches is comparable to variability within batches. This knowledge alleviates the need to always process control and treatment samples together. Finally, in our paper summarizing these analyses (published in *mSystems*[5]) we discuss trade-offs between the number of fractions and replication, and quantify the number of replicates necessary to achieve a given statistical power and detection limit (**Fig 1**).

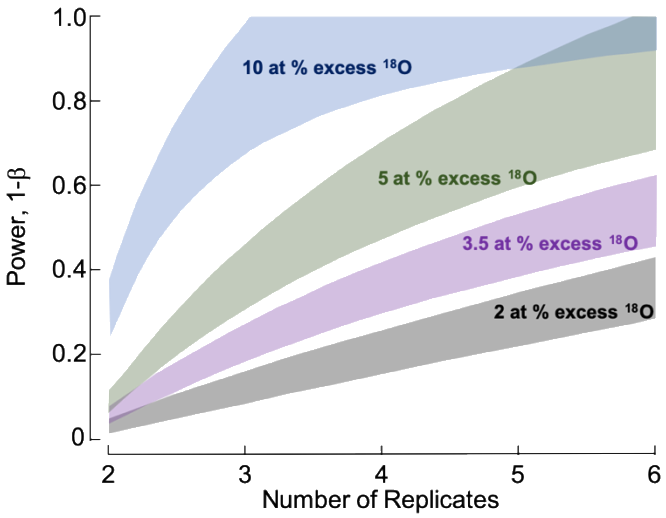

Figure 1. Statistical power of enrichment detection as a function of the number of sample replicates used in a qSIP experiment. To determine the necessary replication for a qSIP experiment, users can choose their desired statistical power and desired detection threshold (represented by different colors).

Normalizing amplicon SIP data

In our calculations of isotope incorporation using SIP-fractionated amplicon counts, we have found evidence of spin and sample artifacts, where the densities of an amplicon have slight variations and need to be corrected. For example, in a ^{13}C-tracer study, we expect the density fraction of a certain amplicon to be constant between the control (^{12}C) samples, and to be the same or denser in

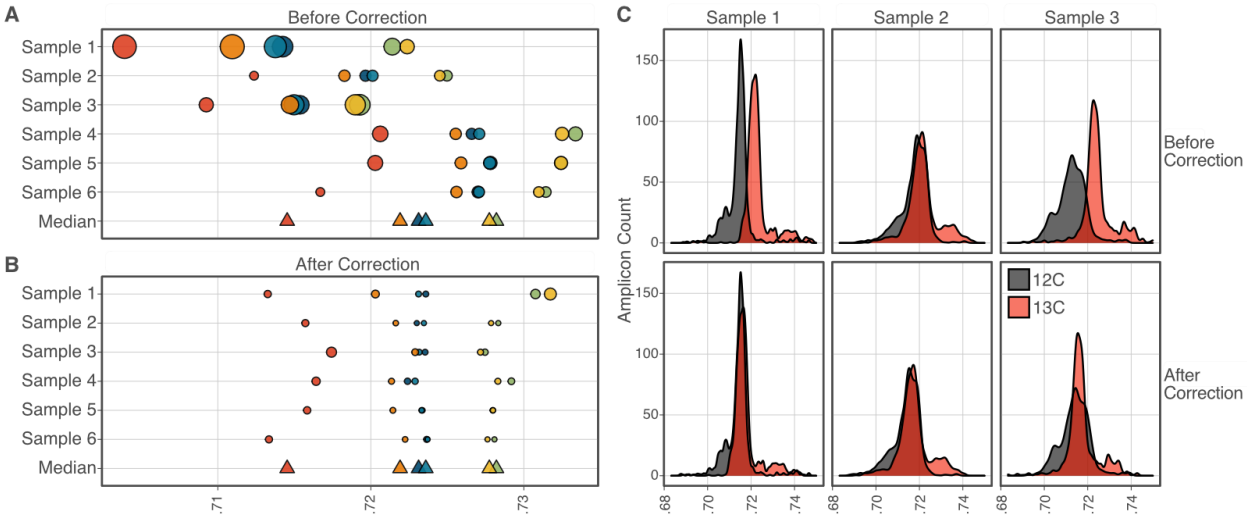

Figure 2. Six amplicons (circles) are shown for 6 samples before (A) and after (B) stress correction. The median value for each amplicon is calculated across the samples (triangles), and a stress value (circle size) is given as the square of the distance from the median. A linear correction is applied to adjust all amplicons within a sample to minimize the stress values. Density curves for full datasets are shown in panel C where the peak of the ^{12}C and ^{13}C better overlap after correction, revealing a 'shoulder' of ^{13}C-enriched amplicons on the right-hand side.

the ^{13}C samples, depending on the amount of heavy isotope incorporation. To correct for such discrepancies, we calculate the "expected" density of each amplicon as the median weighted density among all ^{12}C samples (**Fig 2A**). Next, to correct each individual ^{12}C tube, we obtain a "tube stress" by calculating the square of the density difference between the observed and expected density for each amplicon, and apply a linear shift of the densities to minimize the tube stress value (**Fig 2B**), circle sizes are stress values). Correction for the ^{13}C samples is similar, except we use the expected densities from the ^{12}C samples, and only calculate and minimize the stress on the amplicons towards the lower densities. The reasoning for this is that the lower densities are more likely to be non-enriched with the heavy isotope and are expected to have a density similar to that seen in the ^{12}C samples. This correction approach gives more accurate density curves where the peaks of the samples are more aligned, with a slight increase in the right shoulder of the ^{13}C samples representing enriched amplicons (**Fig 2C**).

qSIP fraction resolution optimization

Our SFA team at NAU is conducting *in silico* experiments to assess qSIP optimization, as a follow up to our recently published sensitivity analysis study[5]. The main purpose of these experiments is to assess fractionation schemes, and how different approaches affect resolution (operationally defined as the standard deviation of the estimate of atom fraction excess tracer content, e.g., AFE ^{18}O). In one set of experiments, we are testing whether the quantity and value of information contained in qSIP fractions is proportional to the amount of DNA they contain, and thus that combining fractions at the tails of the distribution could reduce the cost of the technique with little cost in resolution (**Fig 3**). In contrast, we may find that combining fractions near the center of the peak will cause a substantially larger loss in resolution. We postulate that combining fractions on the left tails (lighter density region) of the distribution will have the least cost in resolution, whereas fractions on the right (higher density region), even though they contain little DNA, hold more valuable information because the right tail is where differences in isotope composition most strongly affect density. The left tail, by contrast, is bounded primarily by taxa GC content.

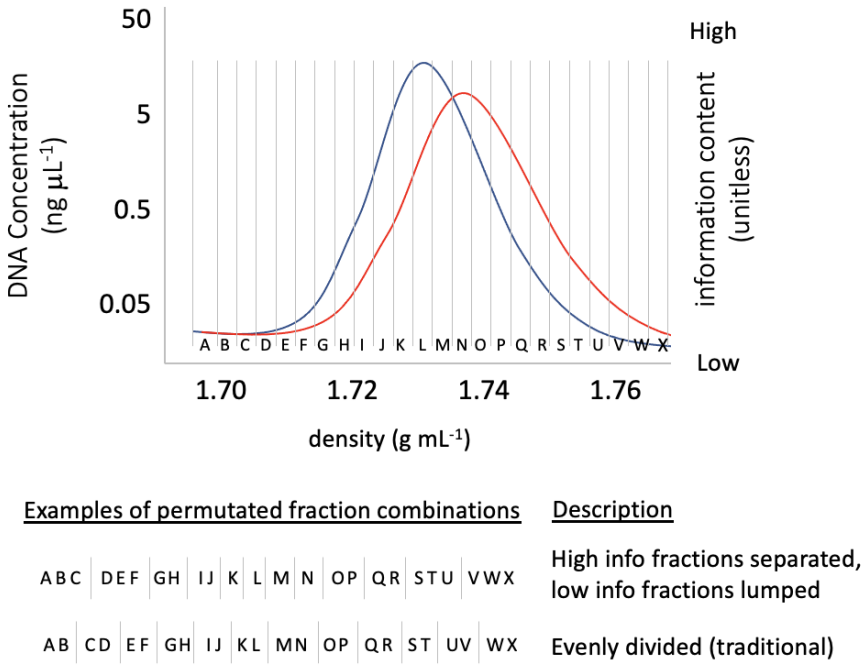

Figure 3. Conceptual scheme for fractionation permutation experiment, a follow-up analysis to Sieradzki et al. 2020. Here, we are testing how combining factions affects resolution in qSIP experiments. Our hypothesis is that schemes that combine fractions with low DNA content will have smaller effects on qSIP resolution, particularly on the left (low density) side of the distribution. The fractions with high DNA concentrations contain more information about taxa occupying a particular density range, and so analyzing these as separate fractions will optimize resolution. Our permutation experiments are examining these effects across multiple ecosystems and conditions, to provide general guidelines for qSIP experiments and *in silico* analyses.

Genome Assembly and Annotation

We use genome-resolved metagenomics to identify ecophysiological traits of populations linked to soil C persistence. New informatics tools have accelerated soil genome-based metagenomic (and metatranscriptomic) analyses[7-9]. It is now possible to assemble large datasets from dozens of samples and recover many 100's of draft quality genomes[10]. Our SFA is developing several tools to facilitate better metagenome curation and viral sequence analyses.

New metagenome assembly methods

One of the benefits of SIP-metagenomes is the increased sequencing depth we achieve, due to the individual sequencing of multiple high-resolution fractions. This increase in sequencing depth helps obtain reads for organisms that would normally be below the limit of detection. However, the increase in sequencing depth makes co-assemblies computationally difficult using traditional metagenomic assembly methods. In collaboration with the Joint Genome Institute, we used one of our soil datasets to co-assemble 95 short read samples (>22 billion total reads, >3.4 Tbp) at once using MetaHipMer[11], producing an assembly of >75 Gbp. As far as we know, this is the largest metagenome assembly to date. This single co-assembly has multiple benefits. 1) Simplicity: our previous co-assemblies of this same dataset had to be processed in 23 batches which makes data management, merging and comparisons between assemblies more difficult. 2) Timing: the CPU time required for MetaHipMer to co-assemble this dataset was 50x faster than even one of our 23 co-assembly batches. This is remarkable given that metagenome assembly requires specialized high-memory machines and their limited supply means 23 co-assemblies usually cannot be conducted in parallel. 3) Quality: the resulting contigs from the MetaHipMer co-assembly are of much higher quality, with an L50 twice as high as the average from the 23 co-assemblies, and with 10x more data present in large contigs (>50Kb). 4) Unified set of contigs: having all of the data in a unified set of contigs has many advantages, notably the increased read-depth of rare organisms and the removal of the need for dereplication steps in metagenome assembled genome (MAG) curation. Indeed, the count, quality and diversity of recovered MAGs increases as more read sets are used (**Fig 4**). Some MAGs present in low-abundance were only recovered from the co-assembly using all read sets, and are likely too low abundance to be assembled and binned with other assembly workflows.

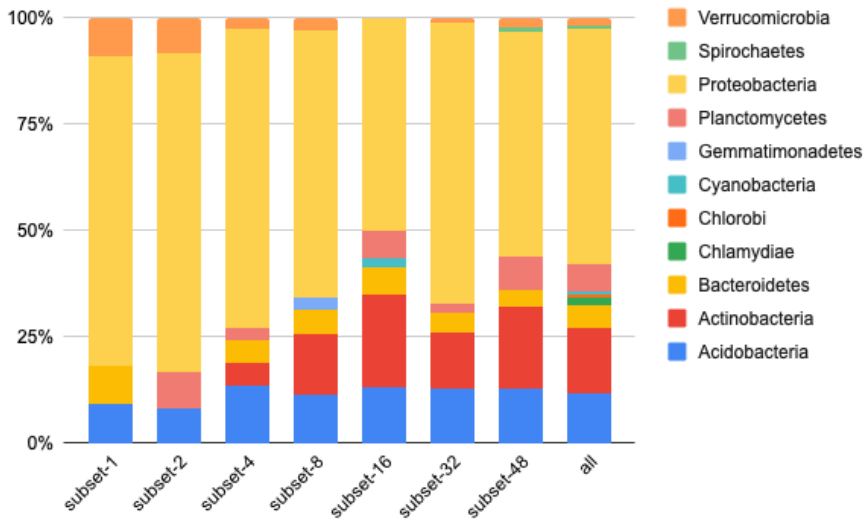

Figure 4. Relative abundance of different microbial taxa with increasing number of read sets used in a MetaHipMer co-assembly. The x-axis represents different grouping strategies for co-assembling the 95 metagenome samples, from 1, 2, 4, 8, 16, 32, 48 to all 95 samples. Taxonomic diversity increases from approx. 4 phyla in the first subset to >10 phyla in the full 8 TB co-assembly. Larger co-assemblies appear to allow us to detect more phylogenetic diversity, including possibly low-abundance microbes.

FixAME: automatic curation and improved metagenomic assembly

High-throughput recovery of MAGs is increasingly one of the primary ways that natural and

experimental microbial communities are characterized. Consequently, the recovery of genomes that accurately reflect true biological entities is essential. Contemporary metagenomic projects, especially those being developed by our Soil Microbiome SFA, can produce thousands of genomes, and the computational assemblies that comprise these mass-produced MAGs contain characteristic errors. Assembly errors perturb or even preclude functional predictions and accurate phylogenetic analyses, and can confound biochemical studies, limiting the full potential utility of MAGs. To resolve these issues, we have informatically assessed the prevalence of assembly errors using three commonly used metagenomic assemblers (MEGAHIT, metaSPAdes, and IDBA-UD) across five environments: soil (from our SFA), lake surface waters[12], surface[13] and deepocean waters[14], and the human gut[15] (**Fig 5**). Assembly errors were found across all tested assemblers and environments, and lower coverage generally resulted in more errors. Assembly errors can be repaired manually, but curation is time-consuming and requires human-guided curation, and so it

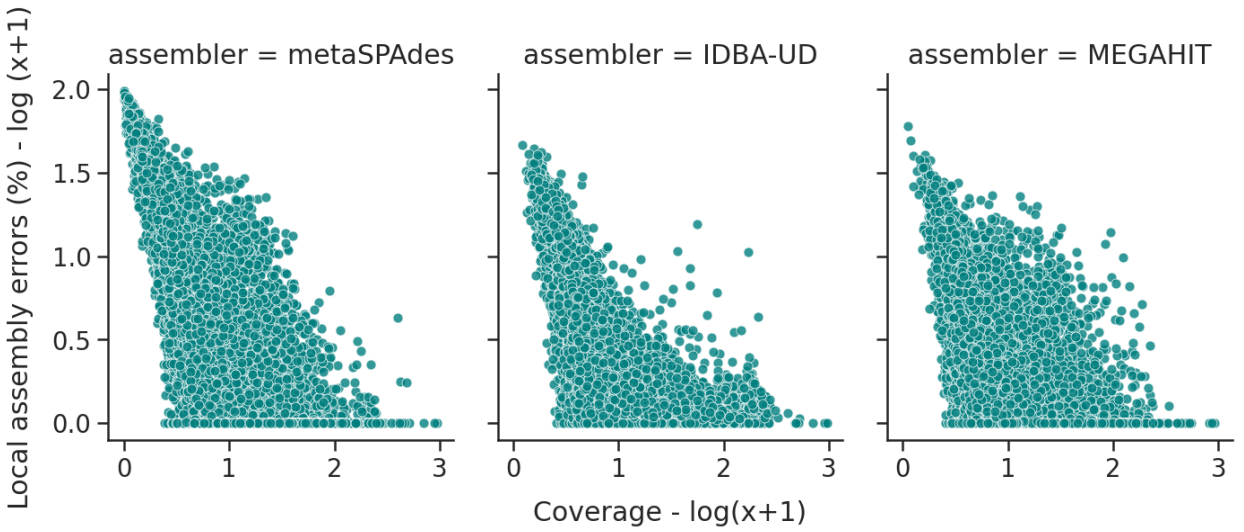

Figure 5. Quantification of assembly errors using MEGAHIT, metaSPAdes, and IDBA-UD across five environments: soil, lake surface waters, surface and deep ocean, and the human gut. Each point represents an assembled sequence >5 kb and the % of bases of that are erroneous versus coverage.

is rarely performed. To overcome this bottleneck, we are developing FixAME, a software toolkit for the automatic curation and improvement of metagenomic assemblies that does not necessitate human intervention. FixAME is not limited to a few genomes and can be run on thousands of genomes or the entire set of assembled sequences from a metagenome. FixAME is being integrated KBase (kbase.us) as a public resource for the easy and efficient improvement of large numbers of genomes by the scientific community. Following the full development of FixAME, we will be able to scale up to curate and improve assemblies in the thousands of MAGs in public databases.

New hybrid long-reads viromics

Assembling virus genomes/fragments to characterize mixed virus communities using short read is a robust method that has enabled diverse ecological insights into the ecosystem impacts of viruses. However, highly variable regions within virus genomes can obscure genome diversity signals, particularly at the strain level, where gene sequence variation could offer insights into biotic and abiotic evolutionary pressures on these genomes. To better capture ‘intra-genome’ diversity (microdiversity), we developed a wet lab and informatics workflow that leverages long-reads to enhance our assembly capabilities[16]. For the informatics workflow (**Fig 6A**), we have benchmarked available tools to QC, error-correct and assemble virus long-reads, as well as use the

SPAdes assembler to perform hybrid assembly. All contigs are then subsequently combined and dereplicated. Using this workflow, we have shown significant improvements in genome size (**Fig 6B**), completeness (**Fig 6C**) and microdiversity (**Fig 6D**) metrics compared to short read-based viromes and our previous VirION method.

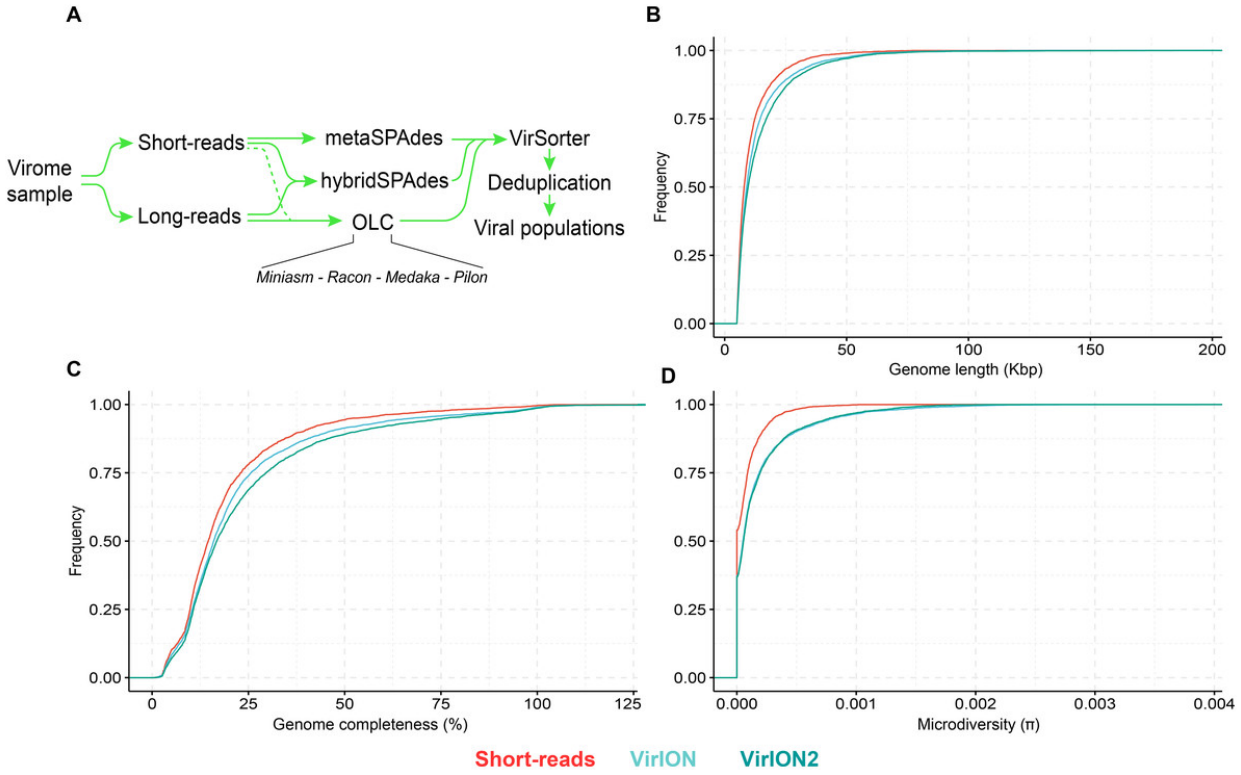

Figure 6. VirION2 informatics pipeline and comparison of virus genome properties between short-read and 'long-read-enhanced' viromes. (A) Workflow to produce 'enhanced viromes', in which Spades, hybrid and long-read (OLC) viruses are combined to maximize the recovery of virus signals. (B) Cumulative Distribution Function (CDF) plot depicting the frequency (y-axis) of virus genomes according to genome length (measured in kilo basepairs (kbp), x-axis) between three assembly strategies. (C) Cumulative Distribution Function (CDF) plot depicting the frequency (y-axis) of virus genomes according to genome 'completeness' (measured in %, x-axis) between three assembly strategies. (D) Cumulative Distribution Function (CDF) plot depicting the frequency (y-axis) of virus genomes according to genome microdiversity per genome (measured as π, x-axis) between three assembly strategies.

iVirus tools on KBase

Studies of environmental viruses, and their influence on mortality, gene transfer and metabolic reprogramming are currently limited by existing informatics tools. Our project has worked to democratize the existing "iVirus" analysis suite by implementing its core components on DOE's KnowledgeBase (KBase) and to develop new analytical tools that enable better host prediction for newly discovered viruses. We have successfully ported several iVirus apps from the CyVerse Cyberinfrastructure, including; a virus identification tool (VirSorter[17]), viral classification (vConTACT2[18]), and a virus-host prediction tool based on a new analytical framework (VirMatcher, https://bitbucket.org/MAVERICLab/virmatcher/) into KBase. This virus-host prediction tool aggregates several existing virus-host methodologies and uses a probabilistic scoring framework to generate a confidence score. Taken together, these apps form a complete, viral ecogenomics toolkit and are available as a public KBase narrative (https://kbase.us/n/75811/85/). Updates to these KBase-enabled iVirus tools have been

concomitant with other, recently introduced, virus-focused tools within KBase, such as DRAM-v, which provides viral annotation and identifies auxiliary metabolic genes[19]. Together, these tools allow generators of environmental metagenome and virome datasets to produce far richer, more contextualized analyses.

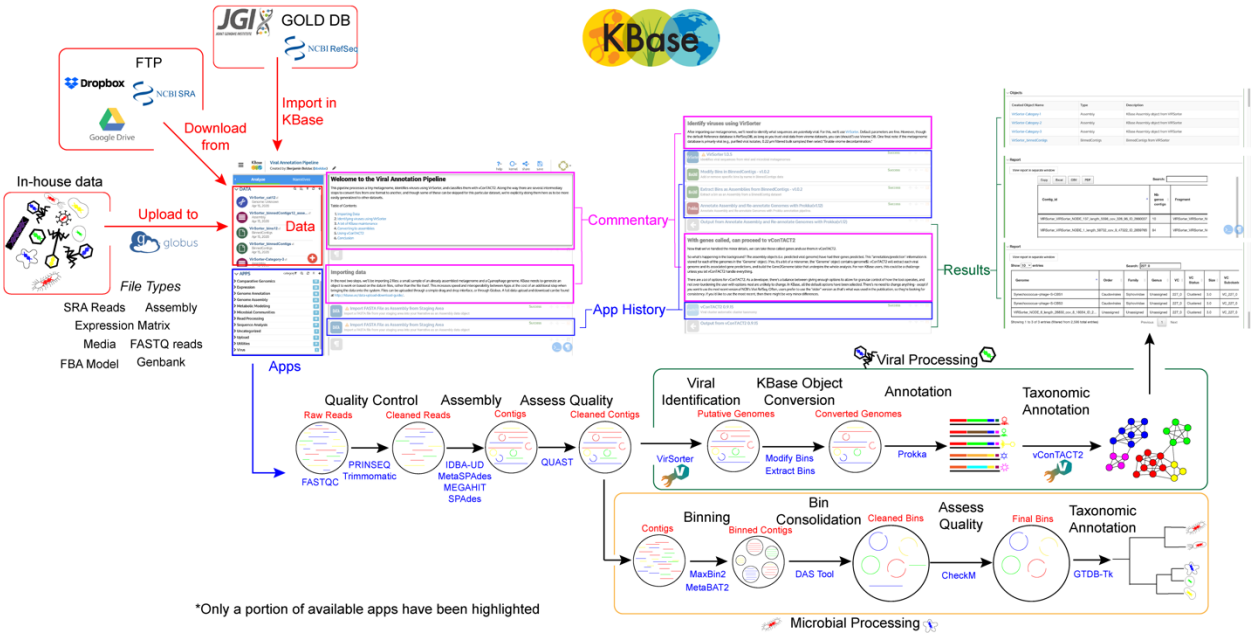

Figure 7. Overview of the viral ecogenomics workflow in KBase supported by the LLNL SoilSFA team. Data sources (outlined in red) can be provided by the user or retrieved from public endpoints. A list of apps (outlined in blue) is searchable and filterable, with "virus" as one such filter in order to quickly find virus-focused tools. User annotations, or notes (outlined in purple) are provided within the Narrative as a means of providing context and background to the analyses. Finally, results (outlined in green) display data generated by the apps. Below is the pipeline where a user can process a viral dataset from raw reads to QC and assembly, viral identification and cleanup, taxonomic annotation, and matching virus-host pairs (not pictured). KBase-powered iVirus apps integrate with existing KBase apps for a complete pipeline, allowing KBase users the option of selecting different apps for the different stages of processing (e.g. different assemblers, quality control, microbial binning tools).

Phanotate: virus gene calling software

Several methods have been developed to identify open reading frames (ORFs) from bacterial genomes, and these methods are also typically used to identify ORFs in virus/phage genomes as well. Phage genomes, however, have certain complexities that bacterial ORF finders do not consider including 1) an extremely high coding density, 2) a higher frequency of overlapping genes, and 3) more instances of ORFs contained entirely within other ORFs. Therefore, these phage-specific genome structural features are often missed by bacterial-specific gene callers. With collaborators at SDSU, viral genome experts at LLNL developed Phanotate[20], the first gene caller specifically designed for viral/phage genomes. Phanotate makes a weighted graph representation of possible

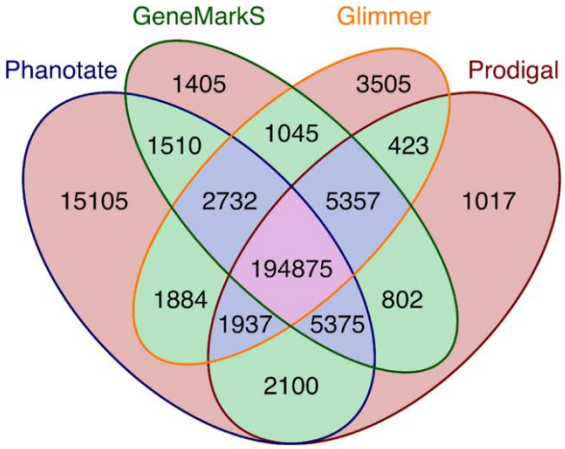

Figure 8. Overlap of predicted gene calls of 2,133 viral genomes from 3 gene callers developed for prokaryotes (GeneMarkS, Glimmer and Prodigal) and one designed for viruses (Phanotate). Many ORFs (82%) were found by all gene callers, however, Phanotate had the most uniquely identified ORFs.

ORFs, to maximize the optimal path through the genome while allowing for overlaps and nested genes. Phanotate was validated against 2,133 phage genomes in NCBI, and compared with results from three popular bacterial gene finders Glimmer[21], GeneMarkS[22], and Prodigal[23]. 239,072 total ORFs were identified among the 4 methods, and there was an agreement in 82% of those gene calls (**Fig 8**). Phanotate found over 15,000 additional ORFs missed by the bacteria-centric methods. Computational validation of these new gene models is difficult – proteomics would be ideal however public peptide databases typically only report peptides matching predicted gene models. Instead, Phanotate gene calls were compared against short reads from >94,000 public metagenomes and validated as likely ORFs due to their higher than expected sequence conservation among these diverse datasets, indicating they are under selective pressures, while other phage regions not predicted to encode protein were less likely to be conserved.

PhATE/MultiPhATE virus annotation pipeline

With the ever-increasing volume of phage genomes being generated from high-throughput sequencing data, there is a need to more rapidly annotate these genomes and make meaningful comparisons of the results. Our SFA viromics team has developed the MultiPhATE[24] annotation pipeline, which structurally and functionally annotates phage genomes using public and/or custom databases, and have included comparative genomics tools to analyze the annotations.

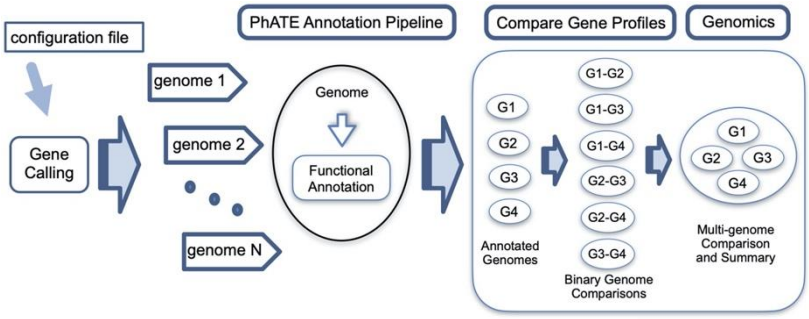

Figure 9. The MultiPhATE workflow begins with genome/contig nucleotide sequences. It proceeds through gene-calling and functional annotation using several custom or public databases. Finally, a comparative genomics workflow is included to compare functional annotations from multiple phage genomes.

The MultiPhATE workflow allows a used to select one or several gene calling algorithms (including Phanotate for virus-tuned gene calls), and a priority or consensus set of gene calls can be retained (**Fig 9**). Next, these gene calls are compared against several functional databases including NCBI, VOGs/pVOGs, Swiss-Prot, etc using both nucleotide (blastn) and amino acid (blastp, HMMer suite) tools. Recently, we release an updated version, MultiPhATE2[25], with several improvements over the original algorithm. These include more options for parallelization to rapidly annotate large collections of viral genomes. The new workflow also enables the discovery of auxiliary metabolic genes (AMGs) with databases such as Carbohydrate-active enzymes database (CAZy), National Center for Biotechnology information protein database (NCBI NR), and Swiss-Prot.

Trait-Based Modeling

One of the goals of the LLNL Soil Microbiome SFA is to build and use a trait-based model (TBM) that evaluates links between ecophysiology and soil C dynamics by combining the recently developed processing scaling theory SUPECA[26] and Dynamic Energy Budget (DEB) theory[27]. This model will allow prediction of biophysical, metabolic and life history traits of bacteria and archaea and their representation in a consistent and theory-based modeling framework. It will also help to identify key fitness traits at the genome or community level and allow model-based hypothesis testing and generation in a reproducible manner.

Bacterial growth efficiency as a species trait

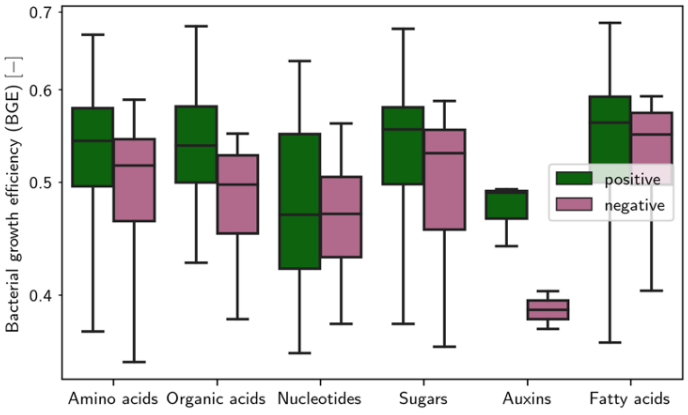

Figure 10. DEB predictions of bacterial growth efficiency for isolates (positive responder = rhizosphere isolate, negative responder = bulk soil isolate) and root exudate compounds, grouped into six classes.

Bacterial growth efficiency (BGE) is the amount of carbon incorporated into biomass versus carbon respired to the atmosphere and it reflects dynamic allocation of a microbe's energy budget to growth under given thermodynamic constraints. BGE is an important parameter in ecosystem models and controller of carbon decomposition in soil. Its central role in pathways of mineral-associated organic matter formation has been postulated for distinct soil compartments, e.g. in the rhizosphere, formation should primarily occur through an *in vivo* microbial turnover pathway and favor carbon substrates that are first biosynthesized with high carbon-use efficiency. In order to understand variation in BGE, we are using the dynamic energy budget (DEB) theory to predict BGE, which allows for a thermodynamically consistent treatment of the balance between structural maintenance, structural growth and extracellular enzyme production in microbial metabolism. We used microTrait (a genomes-to-traits workflow), allometric scaling theory and biophysical modeling to constrain DEB parameters for substrate uptake, assimilation efficiency, depolymerization rates and enzyme allocation, protein synthesis, and maintenance rates. We then conducted batch simulations for 39 bacterial isolates individually grow on 84 root exudate compounds from the wild oat grass *Avena barbata*. We found a significant association between rhizosphere response group and BGE; the BGE of rhizosphere-adapted bacteria was consistently higher across substrate classes (**Fig 10**). DEB predicts a substantial amount of variation in BGE, both at broad (class, ~20%) and fine (strain, ~40%) taxonomic levels. While resource type was a weak predictor across species (~6%), it explained ~50% of variation in BGE within species. Our study suggests that genome-level information together with dynamic energy budget trait-based modeling can resolve variations in BGE within and across microbial communities that should be considered in ecosystem models.

Ohm's law applied to microbial biogeochemistry

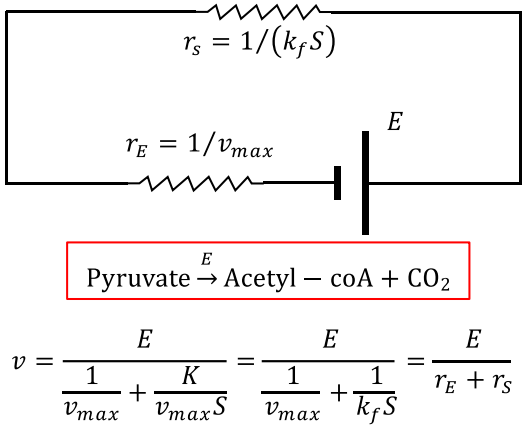

Figure 12. Monod kinetics can be interpreted using Ohm's law. Example in red box depicts the enzymatic decomposition of pyruvate into acetyl-coA and CO_2.

A central challenge in modeling microorganisms and the biogeochemical process they carry out in complex systems such as soil is to represent diverse metabolic pathways and their interactions. Traditionally, this is achieved through the application of Monod kinetics, or the law of mass action. The use of Monod kinetics is simple, but comes with the risk of scaling inconsistency among parameters when increasing from single to many metabolic pathways. In contrast, the law of mass action is much more coherent and rigorous, but is mathematically very difficult for upscaling as the number of metabolic pathways and microbial

populations increase. Our analyses indicate that the Monod kinetics can be interpreted using the analogy of Ohm's law, where, for the case of non-limiting substrate, reaction velocity v is analogous to an electric current that is equal to enzyme abundance E (i.e., the "symbolic" voltage) divided by the mean first passage time (i.e., the "symbolic" resistance) that is defined by enzyme trait r_E (the inverse of maximum reaction rate v_{max}) and the inverse of substrate delivery rate $1/k_f S$ (**Figure 11**). The Ohm's law analogy enables an intuitive yet clear understanding of the physical meanings of substrate kinetics parameters. Its extension to a chain of enzymes enables us to quickly infer the tradeoffs between the length and catalytic speed of a metabolic pathway, and its bioenergetic efficiency (i.e. the amount of Gibbs free energy extracted versus the amount of Gibbs free energy that can be released upon the full oxidation of the substrate): in other words, the longer the metabolic pathway, the slower its metabolic catalysis rate and the higher its bioenergetic efficiency. Additionally, combining Ohm's law with thermodynamics allows us to derive a more comprehensive representation of temperature sensitivity of enzyme-catalyzed reactions. This new representation shows that temperature sensitivity of a metabolic pathway consists of (1) kinetic temperature sensitivity modulated by substrate availability, (2) thermodynamic temperature sensitivity of the chemical reaction of interest, (3) transition state temperature sensitivity, and (4) enzyme conformation temperature sensitivity. In comparison, the more standard (and commonly used) macromolecular rate theory only accounts for components (3) and (4) of our new mathematical representation. When applied to facultative anaerobes, our Ohm's law analogy shows why fermentation is preferred over aerobic respiration under high glucose supply, therefore successfully explaining the Warburg and Crabtree effect[28] observed in biological systems.

Summary

The LLNL *Microbes Persist* Soil Microbiome SFA uses a multi-domain approach to identify the microbial and viral inhabitants of soil ecosystems, providing 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 pushing the boundaries of genome resolved metagenomics, viromics, quantitative stable isotope probing, and trait-based modeling – four powerful and highly synergistic computational approaches. This allows us to make connections between genomically resolved traits, activity, and carbon transformation, giving us an unprecedented picture of the most relevant traits and taxa in soil ecosystems.

Currently, our ability to analyze microbial activities in environmental microbiomes is strongly data limited, for two reasons: 1) there are very few techniques that link performance to genotype in nature, and 2) those that are available – like qSIP – are not yet fully standardized procedures that support rigorous cross-site comparisons and comparison to model output. Our efforts to test the sensitivity, precision, and statistical power of qSIP will allow it to be used more broadly, and with lower costs. New metagenomics computation and curation approaches, such as MetaHipMer and FixAME will streamline assembly and interpretation of large numbers of metagenome assembled genomes. Tools in the iVirus package, that we have recently incorporated into BER's KBase, are also making it far more tractable to extract ecological patterns of viral sequences from complex environmental datasets. Finally, incorporating metrics of microbial function, such as the population metrics measured by qSIP, will improve biogeochemical models. Trait-based modeling is a promising way to integrate information at the genome level (e.g. minimum generation times, substrate utilization capacity, transport kinetics, biomass chemistry, and phage covariance, etc.) to predict emergent processes like total microbial biomass, community composition, turnover and respiration in a way that can dynamically scale from 'omics data to system-level fluxes.

References

1. Murrell, J.C. and A.S. Whiteley, *Stable Isotope Probing and Related Technologies*, ed. J.C. Murrell and A.S. Whiteley. 2011, Washington DC: ASM Press.
2. Hungate, B.A., R.L. Mau, E. Schwartz, J.G. Caporaso, P. Dijkstra, N. Van Gestel, B.J. Koch, C.M. Liu, T.A. McHugh, J.C. Marks, E. Morrissey and L.B. Price, *Quantitative Microbial Ecology Through Stable Isotope Probing.* Applied and Environmental Microbiology, 2015.
3. Koch, B.J., T.A. McHugh, M. Hayer, E. Schwartz, S.J. Blazewicz, P. Dijkstra, N. van Gestel, J.C. Marks, R.L. Mau, E.M. Morrissey, J. Pett-Ridge and B.A. Hungate, *Estimating taxon-specific population dynamics in diverse microbial communities.* Ecosphere, 2018. **9**(1): p. e02090-15.
4. Blazewicz, S.J., B.A. Hungate, B.J. Koch, E.E. Nuccio, E. Morrissey, E.L. Brodie, E. Schwartz, J. Pett-Ridge and M.K. Firestone, *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: p. 1-13.
5. Sieradzki, E.T., B.J. Koch, A. Greenlon, R. Sachdeva, R.R. Malmstrom, R.L. Mau, S.J. Blazewicz, M.K. Firestone, K. Hofmockel, E. Schwartz, B.A. Hungate and J. Pett-Ridge, *Measurement error and resolution in quantitative stable isotope probing: implications for experimental design,* mSystems, 2020. **5**: p. e00151-20. https://doi.org/10.1128/mSystems.00151-20.
7. Starr, E., S. Shi, S. Blazewicz, B.J. Koch, A. Probst, B.A. Hungate, J. Pett-Ridge, M. Firestone and J. Banfield, *Stable isotope informed genome-resolved metagenomics uncovers potential trophic interactions in rhizosphere soil.* bioRxiv (mSystems, in press), 2021. doi.org/10.1101/2020.08.21.262063.
8. Nuccio, E.E., E. Starr, U. Karaoz, E.L. Brodie, J. Zhou, S.G. Tringe, R.R. Malmstrom, T. Woyke, J.F. Banfield, M.K. Firestone and J. Pett-Ridge, *Niche differentiation is spatially and temporally regulated in the rhizosphere.* The ISME Journal, 2020.
9. Starr, E.P., E.E. Nuccio, J. Pett-Ridge, J.F. Banfield and M.K. Firestone, *Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil.* Proceedings of the National Academy of Sciences, 2019. **116**(51): p. 25900-25908.
10. Diamond, S., P.F. Andeer, Z. Li, A. Crits-Christoph, D. Burstein, K. Anantharaman, K.R. Lane, B.C. Thomas, C. Pan and T.R. Northen, *Mediterranean grassland soil C–N compound turnover is dependent on rainfall and depth, and is mediated by genomically divergent microorganisms.* Nature microbiology, 2019. **4**(8): p. 1356-1367.
11. Hofmeyr, S., R. Egan, E. Georganas, A.C. Copeland, R. Riley, A. Clum, E. Eloe-Fadrosh, S. Roux, E. Goltsman and A. Buluç, *Terabase-scale metagenome coassembly with metahipmer.* Scientific reports, 2020. **10**(1): p. 1-11.
12. Garcia, S.L., S.L. Stevens, B. Crary, M. Martinez-Garcia, R. Stepanauskas, T. Woyke, S.G. Tringe, S.G. Andersson, S. Bertilsson and R.R. Malmstrom, *Contrasting patterns of genome-level diversity across distinct co-occurring bacterial populations.* The ISME journal, 2018. **12**(3): p. 742-755.
13. Pesant, S., F. Not, M. Picheral, S. Kandels-Lewis, N. Le Bescot, G. Gorsky, D. Iudicone, E. Karsenti, S. Speich and R. Troublé, *Open science resources for the discovery and analysis of Tara Oceans data.* Scientific data, 2015. **2**(1): p. 1-16.
14. Sachdeva, R., B.J. Campbell and J.F. Heidelberg, *Rare microbes from diverse Earth biomes dominate community activity.* bioRxiv, 2019: p. 636373.
15. Franzosa, E.A., X.C. Morgan, N. Segata, L. Waldron, J. Reyes, A.M. Earl, G. Giannoukos, M.R. Boylan, D. Ciulla and D. Gevers, *Relating the metatranscriptome and metagenome of the human gut.* Proceedings of the National Academy of Sciences, 2014. **111**(22): p. E2329-E2338.
16. Zablocki, O., M. Michelsen, M. Burris, N. Solonenko, J. Warwick-Dugdale, R. Ghosh, J. Pett-Ridge, M.B. Sullivan and B. Temperton, *VirION2: a short-and long-read sequencing and informatics workflow to study the genomic diversity of viruses in nature.* PeerJ, 2021. **9**: p. e11088.
17. Roux, S., F. Enault, B.L. Hurwitz and M.B. Sullivan, *VirSorter: mining viral signal from microbial genomic data.* PeerJ, 2015. **3**: p. e985.

18. Bin Jang, H., B. Bolduc, O. Zablocki, J.H. Kuhn, S. Roux, E.M. Adriaenssens, J.R. Brister, A.M. Kropinski, M. Krupovic, R. Lavigne, D. Turner and M.B. Sullivan, *Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks.* Nature Biotechnology, 2019. 37(6): p. 632-639.
19. Shaffer, M., M.A. Borton, B.B. McGivern, A.A. Zayed, Sabina L. La Rosa, L.M. Solden, P. Liu, A.B. Narrowe, J. Rodríguez-Ramos, B. Bolduc, M.C. Gazitúa, R.A. Daly, G.J. Smith, D.R. Vik, P.B. Pope, M.B. Sullivan, S. Roux, and Kelly C. Wrighton, *DRAM for distilling microbial metabolism to automate the curation of microbiome function.* Nucleic Acids Research, 2020. 48(16): p. 8883-8900.
20. McNair, K., C. Zhou, E.A. Dinsdale and B. Souza, *PHANOTATE: a novel approach to gene identification in phage genomes*, in *Bioinformatics*. 2019. p. 1-6.
21. Ouyang, Z., H. Zhu, J. Wang and Z.-s. She, *Multivariate entropy distance method for prokaryotic gene identification.* Journal of bioinformatics and computational biology, 2004. 2(02): p. 353-373.
22. Besemer, J. and M. Borodovsky, *Heuristic approach to deriving models for gene finding.* Nucleic acids research, 1999. 27(19): p. 3911-3920.
23. Hyatt, D., G.-L. Chen, P. LoCascio, M. Land, F. Larimer and L. Hauser, *Prodigal: prokaryotic gene recognition and translation initiation site identification.* BMC Bioinformatics, 2010. 11(1): p. 119.
24. Ecale Zhou, C.L., S. Malfatti, J. Kimbrel, C. Philipson, K. McNair, T. Hamilton, R. Edwards and B. Souza, *multiPhATE: bioinformatics pipeline for functional annotation of phage isolates.* Bioinformatics, 2019. 35(21): p. 4402-4404.
25. Ecale Zhou, C.L., J. Kimbrel, R. Edwards, K. McNair, B.A. Souza and S. Malfatti, *MultiPhATE2: code for functional annotation and comparison of phage genomes.* G3, 2021. 11(5): p. jkab074.
26. Tang, J. and W.J. Riley, *The SUPECA kinetics for scaling redox reactions in networks of mixed substrates and consumers and an example application to aerobic soil respiration.* Geosci. Model Dev. Discuss., in review, 2017: p. https://doi.org/10.5194/gmd-2017-46, .
27. Kooijman, S., T. Sousa, L. Pecquerie, J. Van der Meer and T. Jager, *From food-dependent statistics to metabolic parameters, a practical guide to the use of dynamic energy budget theory.* Biological Reviews, 2008. 83(4): p. 533-552.
28. Diaz-Ruiz, R., M. Rigoulet and A. Devin, *The Warburg and Crabtree effects: On the origin of cancer cell energy metabolism and of yeast glucose repression.* Biochimica et Biophysica Acta (BBA)-Bioenergetics, 2011. 1807(6): p. 568-576.

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