BSSD 2019 Performance Metric Q4

Goal: Develop metagenomics approaches to assess the functioning of microbial communities in the environment.

Q4 Target: Summarize the latest computational approaches to analyze large complex 'omics' datasets to describe microbial community function in environmental samples.

Summary:

The LANL SFA in Terrestrial Microbial Carbon Cycling aims to inform climate modeling and enable carbon management in terrestrial ecosystems by discovering widespread biological processes that control carbon storage and release in temperate biome soils. To achieve these goals, computational approaches are essential to analyze increasingly complex 'omics' data from microbial communities in environmental samples.

The SFA continues to leverage advances in 'omics measurement technology and computation tools to decipher microbial community function in environmental samples. The computational approach is aligned with the recent shift in the SFA's research approach. When the SFA began 10 years ago, computational approaches principally addressed "who is there" with extremely limited insight into function [1, 2]. Accessible 'omics data for complex soil communities were primarily amplicon DNA sequences of taxonomic or functional marker genes obtained by PCR, cloning, and Sanger sequencing. Computational tools for community analysis were nascent. Over the past decade, genome resources (DOE-Integrated Microbial Genomics database) and data acquisition increased 1000-fold, common computational pipelines for many analysis tasks have arisen, and the variety of 'omics' measurements has expanded to routinely include shotgun metagenomics, metatranscriptomics, and versions of metabolomics. The SFA is capitalizing on these advances while also developing new computational techniques to decipher how communities with different patterns of carbon cycling function from the species to ecosystem level.

The SFA is establishing for routine use a computational approach that integrates exascale metagenomic computing with multi-scale ecosystem modeling. The computational approach exploits machine learning techniques [3] to reduce the dimensionality of larger 'omics datasets that arise from the SFA's research approach. These techniques yield a subset of features that best predict functional outcomes from microbial community activity [3]. Related computational techniques are used to infer interactions between organisms and metabolic products, yielding specific research targets for mechanistic studies. Comparative genomics [4], exascale computing (NERSC user project) for community metagenome assembly, and metabolic interpolation approaches [5] are being applied to improve the quality and functional interpretation of metatranscriptomic data, which is the most detailed measurement available for community physiology. These approaches are embedded in a larger framework of modeling and simulation of soil carbon cycling. For soil carbon modeling, we are using SOMic 1.0, developed by Cornell collaborators [6]. The SFA is applying SOMic at multiple scales (e.g. [7] to guide the design and interpretation of experiments to understand microbial community functional processes influencing terrestrial carbon cycling.

Background

The revolution in technology for 'omics' measurements that began over a decade ago continues to evolve rapidly, providing ever larger and more complex datasets that document the organisms, physiological activity, and metabolic compounds present in microbial communities. Coupled with parallel advances in computational resources and techniques, stronger insights into how microbial communities in environmental samples function are possible. The latest computational approaches in use and development in the SFA are described below in the context of the following high-level questions the approaches address:

- What organisms are present?
- Which organisms are the major players in functional processes of interest?
- What interactions occur that influence function?
- What specific physiological processes contribute to variation in community function?
- What are the consequences of contrasting microbial community functional patterns on ecosystems over larger temporal and spatial scales?

What organisms are present?

When the SFA began 10 years ago, computational approaches principally addressed "who is there" with extremely limited insight into function. Accessible 'omics data for complex communities at that time were primarily amplicon sequences of taxonomic or functional marker genes obtained by PCR, cloning, and Sanger sequencing. The SFA produced the first large scale (10,000 to 20,000 DNA sequences) targeted metagenomic datasets at the time from soil samples from DOE-funded field studies examining the impact of elevated atmospheric CO_2 and other factors in six terrestrial ecosystems [1, 8-10]. Computational tools for community analysis were nascent. Computation involved of a suite of separate software tools for sequence alignment, clustering into operational units, custom software developed within the SFA to de-multiplex samples and create OTU tables, ordination methods to visualize sample similarity, and ecological statistical techniques. As DNA sequencing technology rapidly evolved, the SFA leveraged external advances in computational pipelines for amplicon sequence processing [1, 2, 8-10]. Two dominant platforms currently exist with a suite of integrated algorithms for quality control and rapid processing of vastly larger quantities of data. The two platforms—USEARCH [11] and QIIME [12]—have become community standards, although debate about their relative merits continues. The SFA currently relies on the USEARCH pipeline for amplicon sequence processing but periodically re-assesses the performance of QIIME. For taxonomic identification of the sequence clusters provided by USEARCH, the SFA uses the bacterial classifier and fungal taxonomic classifiers (to which the SFA contributed [13-15])) hosted by the Ribosomal Data Base. This approach is foundational for every experiment in the SFA prior to down-selection of samples for deeper and more costly 'omics studies. The approach accounts for about 2000 bacterial and fungal community profiles currently being processed within the SFA.

Which organisms are the major players in functional processes of interest?

To accelerate progress toward program goals, the SFA recently shifted from cataloguing microbial *responses* to environment change (old research strategy) to identifying microbial *drivers* of ecosystem function (new research strategy). To discover widespread microbial processes that drive variation in carbon cycling, the SFA has implemented a clinical research paradigm. The clinical approach involves screening a large number of soil communities to identify cohorts that represent contrasting patterns of carbon cycling [16]. The use of cohorts of

communities from diverse geographic locations facilitates discovery of cosmopolitan features that drive functional effects. After taxonomic profiling of communities in each cohort by amplicon sequencing and data processing via USEARCH and the RDP, the profiles are used to distinguish a core set of major players that underpin the contrasting patterns of carbon cycling. The SFA developed a new computational approach to achieve this task [3].

The computational approach capitalizes on the greater rigor inherent in supervised machine learning techniques, wherein data are split into separate training and test sets. The approach uses three different methods—neural networks, random forest decision trees, and indicator species analysis—to find and rank microbial features that predict a target variable (e.g. CO_2 or dissolved organic carbon from soil organic matter decomposition) in regression models with the training data. The features are partially validated by applying the regression models to held-out test data (Figure 1).

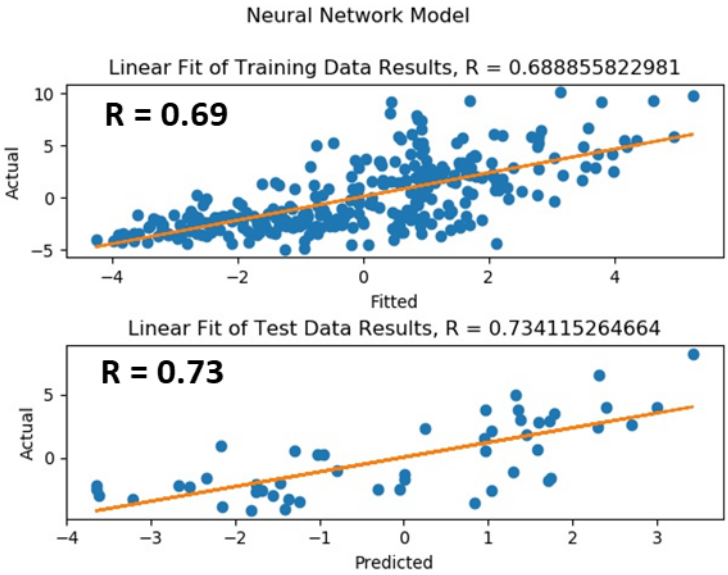

Figure 1. Regression model predicting dissolved organic carbon abundance in microcosms of several hundred microbial communities decomposing plant litter. Top panel shows prediction performance with training data. Features were ranked by a neural network algorithm. Bottom panel shows prediction performance with the same features and regression model applied to test data. From [5].

The top features common to all three methods are down-selected as the most robust features for focused analysis (Figure 2). Down-selecting a core set of 'omics features (in this case, taxa) is useful because it reduces the dimensionality of the data. For example, in analyses thus far, the approach has reduced the complexity of microbiome taxonomic features by more than 20-fold from over 2000 features to about less than 100. The down-selected features provide a more tractable stepping stone to discover mechanisms through inferential analysis of organism traits in published literature [5] or by further experimental approaches. Both avenues are currently being

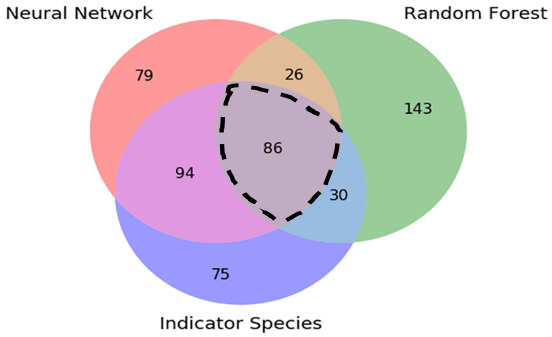

Figure 2. Overlap of taxonomic features from 3 techniques that predict DOC abundance in pine litter microcosms at day 44 of decomposition. The tri-method intersection set (enclosed by dashed line) is the most robust subset of features for further study. From [5].

used in the SFA.

This approach is aligned with the SFA's clinical research paradigm, which provides the much larger quantity of data needed to adequately support use of machine learning techniques. The software is available on GitHub [3] and is already in use by other research groups. Although there are several other platforms (e.g. QIIME2) that facilitate generic use of machine learning techniques with microbial community data, our analysis software provides several unique options that make it more accessible and useful for microbial ecologists. For example, the software facilitates sensitivity analyses that allow users to assess the stability of their results as a function of the number of samples (communities) represented in their dataset. The software also allows users to confirm the subset of top features common to the three methods is an acceptable subset for predicting the target functional variable (Figure 3). The software also allows users to assess how prediction performance increases as the ranked features in the set are sequentially added to a regression model.

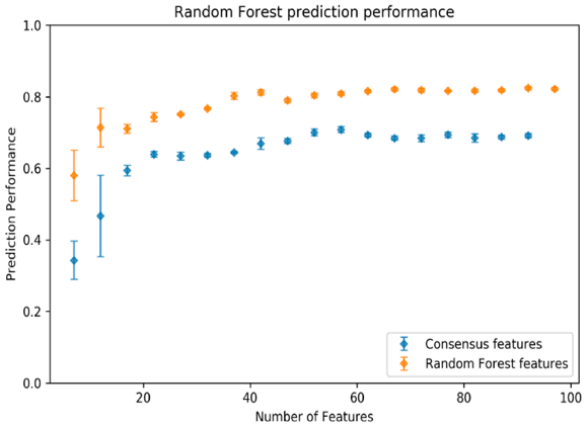

Figure 3. Prediction performance of consensus features. Performance is measured as the Pearson correlation between observed DOC and predicted DOC with the given features in a Random Forest model.

What interactions occur that affect function?

Feature lists from the SFA's machine learning platform are an important first step toward discovering the mechanism(s) maintaining C flow variation, but additional insights about interactions from the molecular to ecosystem level can be obtained from other machine learning techniques. Probabilistic graph modeling is a class of techniques of growing interest across disciplines because it facilitates uncertainty quantification. Within this class of techniques, the SFA began exploring the use of Bayesian networks to assess the structure of interactions among variables of interest (i.e. taxa and DOC). Probabilistic graphical models like Bayesian networks attempt to model the joint probability distribution of a set of random variables in terms of a graphical structure. The nodes of the graph represent random variables, and the edges between nodes represent dependencies between the random variables.

To explore interactions between taxa and dissolved organic carbon (DOC) from soil organic matter decomposition, the SFA first uses our basic machine learning platform [3] to significantly reduce the number of variables. The reduced set of taxa is then used to model the joint distribution of taxa and DOC with a Bayesian network. Once the joint probability distribution between taxa and DOC is learned from training data, the model can be applied to prediction of DOC on held-out testing data. A Bayesian network structure determined from pine litter microcosm communities is shown in Figure 4. Classification accuracy (i.e., predicting DOC abundance) on held-out test data was >80%, which is a promising start. Because this approach is

more computationally intensive, our current implementation limits the number of variables (nodes) to about 20 or less.

A useful aspect of this approach is that it has the potential to predict directionality of interactions. At one extreme, interactions can be strictly one-directional among a suite of taxa with DOC as either the top parent node or the final child node. Such a structure would indicate a hierarchical network of metabolic hand-offs among taxa consuming DOC or producing DOC, respectively. This outcome would suggest that a single organism near the top of the hierarchy could be manipulated to alter DOC abundance. At the other extreme, a non-hierarchical network could occur, suggesting a loose confederation of organisms that work independently in the decomposition process and all taxa must be manipulated in aggregate to alter carbon flow. Distinguishing these contrasting possibilities is important because the outcome strongly shapes discovery and management of microbial mechanisms controlling carbon flow in soils.

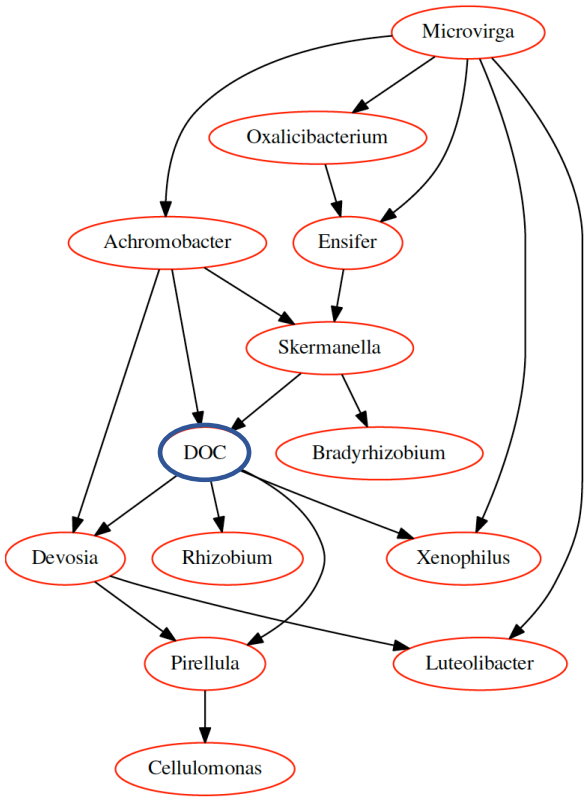

Figure 4. Discrete Bayesian network model of the dependency between DOC and a set of 12 selected genera. Graph structure was inferred from pine litter microbiome data.

With continued development, we expect to obtain network models like Figure 4 that substantially improve interpretation by showing the degree of statistical support for each directional connection in the network. Initial attempts to apply Bayesian network models in a cross-ecosystem validation study are suggesting some aspects of network architecture are conserved for decomposer communities on different types of litter. These encouraging results suggests the combination of the SFA's experimental strategy and the network cross-validation approach may facilitate discovery of robust phenomena among ecosystems. The SFA is continuing to develop this approach. The central challenge is clear representation of uncertainty. Without uncertainty quantification, it is easy to generate irrelevant or misleading network structures. Further development includes assessing performance relative to techniques developed by others [17, 18] and manipulative experiments to validate inferences.

A benefit of using taxonomic profiles as source data to infer community interaction networks is that the data quantity (i.e., number of community profiles) required to construct these statistical models is more tractable. A weakness is that bacterial and fungal components of microbial communities are separately profiled, hindering integrated analysis. The SFA is addressing this limitation by developing a new profiling method that enables integration of bacterial and fungal taxonomic profiles [19].

What specific physiological processes contribute to variation in community function?

Given a set of core taxa and interactions (from methods above) that drive contrasting patterns of carbon flow, the SFA uses three approaches to assess the underlying distinctive physiological processes --- inference from published phenomic data, inference from genomic data, and analysis of metatranscriptome data.

Inference from published phenomic data is a labor-intensive literature search for traits that are reported as characteristic of specific taxa [5]. For example, some taxa are specialists for specific types of metabolism (e.g., photosynthesis, methanotrophy, inorganic sulfur cycling) or ecological strategies (prolific antibiotic production, specialization as oligotrophs, predation of other microbes). After literature mining, routine statistics are applied to assess significant differences between community cohorts [5]. The absence of an appropriate trait database is a resource gap. The SFA is exploring ways to address this gap in concert with external groups.

Inference from reference genomes in DOE's IMG is a computational capability established by others in 2013 [20]. For example, the PICRUSt tool (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States) is a bioinformatics software package that matches a bacterial taxonomic profile (16S rRNA gene sequences) from a community to references genomes and derives a predicted metagenomic functional profile [20]. The SFA has used this software to infer potential physiological activities that significantly differ in abundance between plant litter decomposer community cohorts with contrasting patterns of carbon flow (Figure 5) [5]. Owing to known limitations of the PICRUSt tool, the SFA recapitulates the underlying algorithm manually with the subset of core taxa derived from machine learning techniques described in previous sections above. The SFA is exploring automation and modifications of the process that will improve data interpretation.

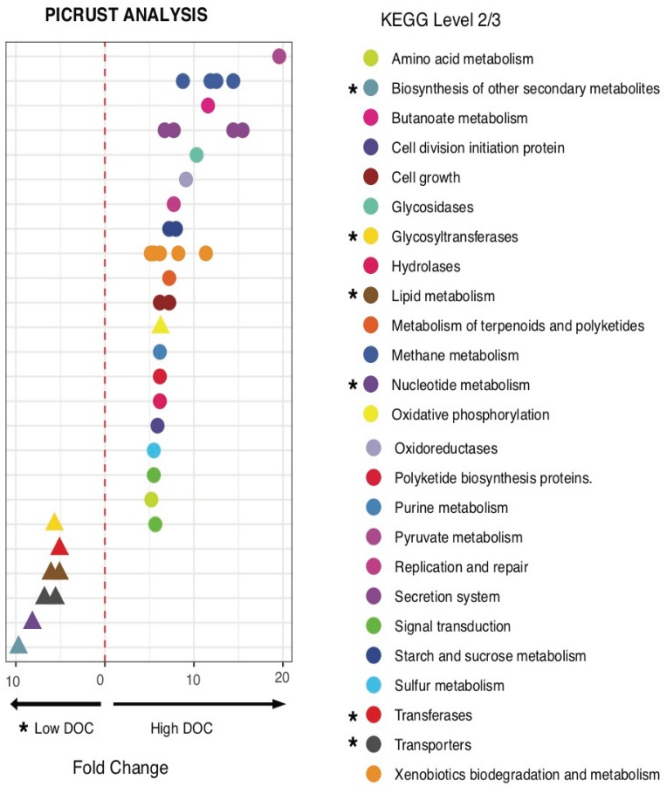

Figure 5. Physiological differences in PICRUSt predicted metagenomes in community cohorts with high versus low abundance in DOC after 44 days of plant litter decomposition in microcosms. Shown are 25 functional processes with significant 5-fold or greater differences in abundance.

Phenomic and genomic inferences generally reflect functional *potential*. To obtain a more complete perspective, these inferences are integrated with analysis of metatranscriptome data, which is a more direct measurement of community activity. The SFA currently uses a combination of MG-RAST (supported by Argonne National Lab) and a custom workflow on our own server as the principal approach for functional annotation and analysis of shotgun metagenomic and metatranscriptomic

data [5]. In an external benchmark study, MG-RAST performed the best for functional analysis of metagenomes [20]. We supplement the MG-RAST analysis with a custom pipeline that enables specific annotation of Carbohydrate Active enZymes (Figure 5). We recently evaluated KBase as a potential platform for metatranscriptome analyses, wherein a suite of applications

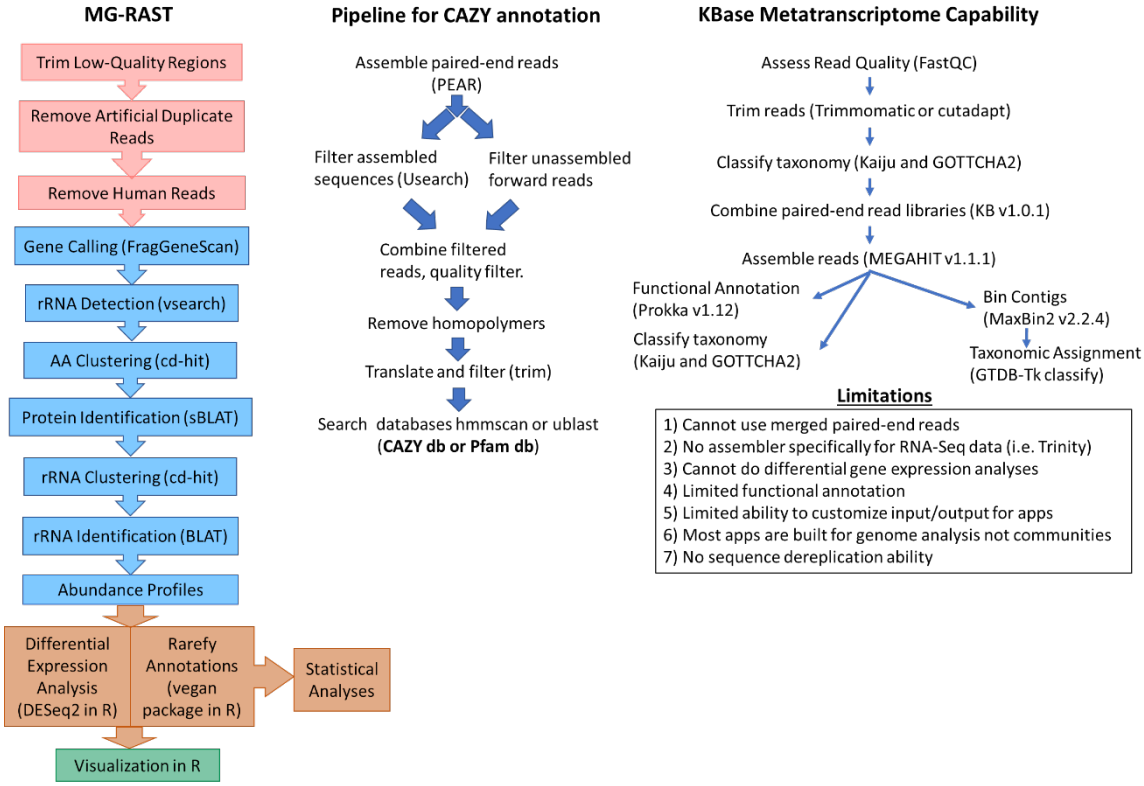

Figure 6. Metatranscriptome analysis pipelines. At present, the SFA routinely uses MG-RAST and separately, an internal server for annotation of CAZymes. A suitable pipeline in KBase is not yet available.

could be strung together provide a complete analysis pipeline. However, metatranscriptome analysis is a nascent capability at present in KBase; a number of analysis gaps will require installation of additional applications on KBase to achieve a complete pipeline (Figure 6).

Exascale computing to improve metatranscriptome analysis. A fundamental step that is routine in transcriptome (single-organism) studies but remains difficult to achieve with soil metatranscriptomics is assembly of reads into contigs. Assembly increases sequence lengths, which is important for greater annotation accuracy, especially for complex microbial communities containing many species that are only distantly related to existing reference genomes. Assembly can also enable detection of differential expression at the gene level (instead of gene-family level) by creating gene scaffolds to which sequence reads can be mapped. Thus far in the SFA, attempts to assemble metagenome or metatranscriptome data have provided so few contigs that analyses revert to the cruder use of unassembled sequence data [21].

As a step toward addressing the assembly problem, the SFA launched an effort through a NERSC user-facility proposal to build ecosystem consensus assemblies that the research community can leverage to map transcriptome reads from a single study. The concept involves gathering all metagenomic and metatranscriptomic data available for a particular ecosystem (or ecosystem type) produced from different experiments, labs, and sequencing platforms. The

collective data increase the likelihood of achieving an effective assembly, which can be sequentially improved over time as additional data become available.

To test the concept, the SFA is collaborating with other LANL scientists (Dr.s Patrick Chain and Migun Shakya) and with the developers of the MetaHipMer assembly program to create a soil biocrust consensus assembly. This is a unique dataset (~ 1TB) to test the ultra-fast and highly scalable assembler, MetaHipMer [22]. The large dataset consists of sequences generated using older 454 sequencing technology and current Illumina sequencers, and both metagenome and metatranscriptomes derived from different biocrusts at different times mostly from prior studies in the SFA. Because it is a diverse dataset, it requires large shared memory and currently starts assembly using 800 nodes (~ 64GB X 800 = 51.2TB) in KNL architecture of Cori. We estimate that it will require at least 300 minutes of wall clock time, which equates to around 15M CPU minutes (250K hours). This effort will also explore the effect of varying k-mer sizes on assembly quality, which will require similar time for assembly. If effort is successful, the SFA will use transition to using this approach as a foundational step to improve the accuracy and insight for future transcriptome studies.

Comparative genomics to refine metatranscriptome analysis. A desired goal with environmental metatranscriptomics is to relate the abundance of expressed metabolic pathways to rates of biogeochemical cycling (e.g. the rate or prevalence of inorganic nitrogen cycling, which is linked to carbon cycling). This goal is undermined in part by errors in quantification of the expression of key pathways. One source of error is failure to account for expression of incomplete pathways. That is, some organisms may contain and express an incomplete set of genes for a pathway, rendering the pathway nonfunctional yet contributing to a false positive metatranscriptomic signal of pathway activity. The SFA began addressing this problem through comparative genomics, leveraging 6000 complete genomes in DOE's IMG database. Incomplete N-cycling pathways were found in 39% of surveyed species [4]. The vast majority of species with complete pathways had limited ability to utilize inorganic N in multiple oxidation states, which suggests the extent of nitrogen leaching from ecosystems is influenced by the composition and spatial organization of species that must perform inter-species nitrogen transfers [4]. The SFA is extending this approach to other elemental cycles and pursuing a metatranscriptomic analysis capability that will automatically account for potential false positives.

What are the consequences of microbial community functional patterns over large temporal and spatial scales?

A distinctive facet of the SFA's recent shift in research strategy has been to use soil carbon modeling and simulation as an overarching framework to guide research priorities, experimental design, and to improve data interpretation. The SFA is using SOMic, a microbially-based SOC model recently developed by Cornell collaborators Dr.s Dominic Woolf and Johannes Lehmann [6]. SOMic captures the recent shift in concepts about organic matter persistence, wherein persistence depends on microbe-mineral-organic matter interactions, instead of recalcitrance of

the organic matter. As described in [6], "SOMic assumes that microorganisms take up only dissolved organic carbon (DOC), because substrates must be in solution to cross the cell membrane (Figure 7). Microbial uptake of DOC competes with sorption to minerals and occlusion within aggregates, whose rate is determined by mineral surface area (approximated by the clay fraction). Microbial uptake is then apportioned between growth and respired CO2 according to microbial C-use efficiency, which is dependent on temperature. Organic matter inputs undergo depolymerization and/or dissolution before entering the DOC pool. Rates of depolymerization and dissolution of organic matter, and desorption of mineral-stabilized SOC are mediated by microbial enzyme activity according to reverse Michaelis-Menten dynamics." Key insights from SOMic simulations suggest SOC persistence may depend on ecological constraints---microbial interactions with mineral-associated carbon---to a greater degree than assumed in most previous models [6]. Based on this insight, the SFA has launched research investigating variation in microbial community features that drive variation in interactions with mineral-associated carbon.

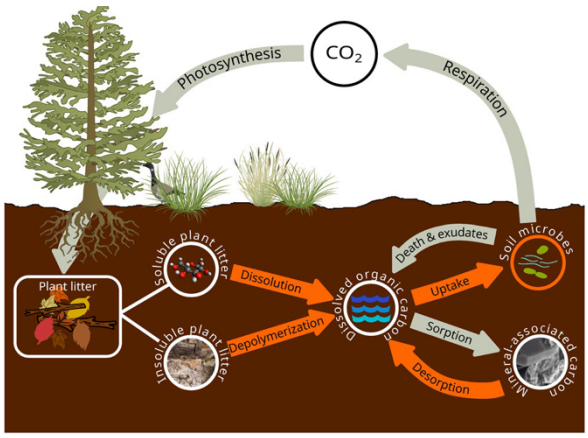

Figure 7. Schematic of the SOMic 1.0 model for soil organic carbon (SOC). From [6].

The SFA is using to SOMic to guide experimental design by simulating potential consequences of microbial-driven variation in carbon flow observed in small-scale experiments at larger spatial and temporal scales. The simulations provide predicted dynamics of microbial biomass, respiration, and soil organic matter abundance. Consequently, the simulations aid planning for the timing of measurements needed to capture key microbial dynamics as well as the time interval required for significant ecosystem changes to be detectable. For example, using the range of differences observed among soil communities decomposing surface plant litter in a prior microcosm study, we applied SOMic to simulate potential consequences in a more complex mesocosm system over a longer time-course with many cycles of litter renewal and decomposition (Figure 8). Five different litter decomposition patterns were modeled. The simulations estimated the minimum number of cycles required for an experiment (currently underway) to achieve detectable differences in soil carbon abundance – a crucial detail for experimental design.

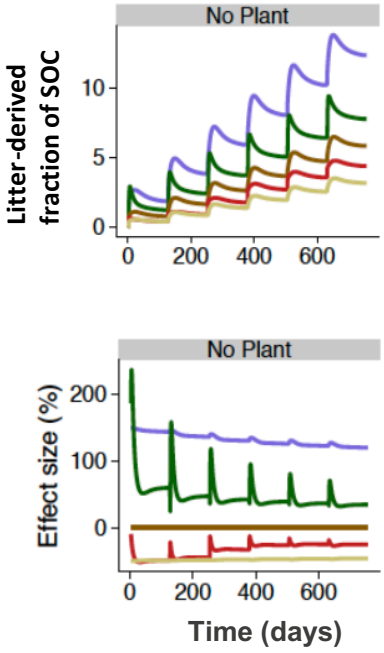

Figure 8. SOMic predictions of soil organic carbon dynamics in mesocosms with different levels of DOC flux.

The SFA is also using SOMic to improve data interpretation [7]. A strength of SOMic is that it can be applied to gain insight into microbial dynamics and soil carbon cycling at many scales,

ranging from days in simple laboratory microcosms [7] to the decades at the global scale [6]. We used SOMic to assess if differences in the quantity of microbial biomass added to laboratory microcosms could be a substantial factor in the large variation in carbon flow observed among communities decomposing plant litter over a short (6-week) timescale [7]. The simulations from SOMic showed respiration (CO_2 efflux) dynamics that matched experimental measurements (Figure 9) and provided substantial information such as probable microbial biomass dynamics that was not possible to measure in the experiment.

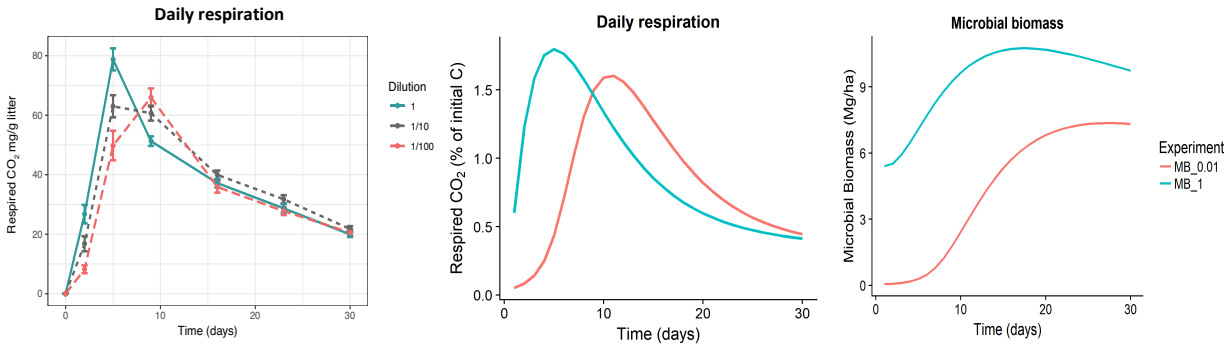

Figure 9. SOMic predictions of microbial respiration and biomass dynamics in plant litter decomposition microcosms initiated with 100-fold differences in initial microbial biomass. Left panel – Measured daily respiration. Middle and right panels – SOMic predictions.

The SFA's new focus on tight integration of modeling (SOMic) and experimentation will enable further refinement of SOMic, which will enhance the power of this capability to increase the quality, efficiency, and pace of the SFA research program.

Bibliography

1. Dunbar, J., S.A. Eichorst, L. Gallegos-Graves, S. Silva, G. Xie, N. Hengartner, B.A. Hungate, R.B. Jackson, D.R. Zak, R. Vilgalys, R.D. Evans, C.W. Schadt, J.P. Megonigal, and C.R. Kuske, *Common bacterial responses in six ecosystems exposed to ten years of elevated atmospheric carbon dioxide.* Environ Microbiol, 2012.
2. Dunbar, J., L. Gallegos-Graves, B. Steven, R. Mueller, C. Hesse, D.R. Zak, and C.R. Kuske, *Surface soil fungal and bacterial communities in aspen stands are resilient to eleven years of elevated CO2 and O-3.* Soil Biology & Biochemistry, 2014. **76**: p. 227-234.
3. Thompson, J., R. Johansen, J. Dunbar, and B. Munsky, *Machine learning to predict microbial community functions: An analysis of dissolved organic carbon from litter decomposition.* PLoS ONE, 2019. **14**: p. e0215502.
4. Albright, M.B.N., B. Timalsina, J. Martiny, and J. Dunbar, *Comparative genomics of nitrogen cycling pathways in bacteria and archaea.* Microb Ecol, 2018. **77**: p. 597-606.
5. Albright, M.B.N., J. Thompson, R. Johansen, D.E.M. Ulrich, L.V. Gallegos-Graves, B. Munsky, and J. Dunbar, *Microbial physiology linked to divergent carbon flow from litter decomposition.* Frontiers in Microbiology, 2019. **Submitted**.
6. Woolf, D. and J. Lehmann, *Microbial models with minimal mineral protection can explain long-term soil organic carbon persistence.* Scientific Reports, 2019. **9**: p. 6522.

7. Albright, M.B.N., A. Runde, D. Lopez, J. Gans, S. Sevanto, D. Woolf, and J. Dunbar, *Initial microbial biomass abundance is a weak driver of variation in CO2 flux during plant litter decomposition.* PloS ONE, 2019. **submitted**.
8. Weber, C.F., D.R. Zak, B.A. Hungate, R.B. Jackson, R. Vilgalys, R.D. Evans, C.W. Schadt, J.P. Megonigal, and C.R. Kuske, *Responses of soil cellulolytic fungal communities to elevated atmospheric CO(2) are complex and variable across five ecosystems.* Environ Microbiol, 2011. **13**(10): p. 2778-93.
9. Berthrong, S.T., C.M. Yeager, L. Gallegos-Graves, B. Steven, S.A. Eichorst, R.B. Jackson, and C.R. Kuske, *Nitrogen fertilization has a stronger effect on soil nitrogen-fixing bacterial communities than elevated atmospheric CO2.* Appl Environ Microbiol, 2014. **80**(10): p. 3103-12.
10. Steven B, L.G.-G., J Belnap, CR Kuske, *Dryland soil bacterial communities display spatial biogeographic patterns associated with soil depth and soil parent material.* FEMS Microbiol Ecol, 2013.
11. Edgars, R.C., *UPARSE: Highly accurate OTU sequences from microbial amplicon reads.* Nature Methods, 2013. **10**: p. 996.
12. Bolyen, E. and et_al., *QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science.* PEERJ, 2018. **https://peerj.com/preprints/27295/**.
13. Liu, K.L., A. Porras-Alfaro, C.R. Kuske, S.A. Eichorst, and G. Xie, *Accurate, rapid taxonomic classification of fungal large-subunit rRNA genes.* Appl Environ Microbiol, 2012. **78**(5): p. 1523-33.
14. Porras-Alfaro, A., K.L. Liu, C.R. Kuske, and G. Xie, *From genus to phylum: large-subunit and internal transcribed spacer rRNA operon regions show similar classification accuracies influenced by database composition.* Appl Environ Microbiol, 2014. **80**(3): p. 829-40.
15. Deshpande, V., Q. Wang, P. Greenfield, M. Charleston, A. Porras-Alfaro, C.R. Kuske, J.R. Cole, D.J. Midgley, and N. Tran-Dinh, *Fungal identification using a Bayesian classifier and the Warcup training set of internal transcribed spacer sequences.* Mycologia, 2016. **108**(1): p. 1-5.
16. Albright, M.B.N., J. Thompson, R. Johansen, D. Lopez, L.V. Gallegos-Graves, A. Runde, R. Mueller, A. Washburne, B. Munsky, T. Yoshida, and J. Dunbar, *Microbial community-level features linked to divergent carbon flows during litter decomposition in a constant environment.* Microbial Ecology, 2019. **submitted**.
17. Basu, S., K. Kumbier, J.B. Brown, and B. Yu, *Iterative random forests to discover predictive and stable high-order interactions.* PNAS, 2018. **115**: p. 1943-1948.
18. Huynh-Thu, V.A., A. irrthum, L. Wehenkel, and P. Geurts, *Inferring Regulatory Networks from Expression Data Using Tree-Based Methods.* PLoS ONE, 2010. **5**: p. e12776.
19. Hutchinson, M., T. Bell, L. Gallegos-Graves, M.B.N. Albright, and J. Dunbar, *Merging fungal and bacterial community profiles via an internal standard.* Journal of Microbiological Methods, 2019. **in prep**.
20. Lindgreen, S., K.L. Adair, and P.P. Gardner, *An evaluation of the accuracy and speed of metagenome analysis tools.* Scientific Reports, 2016. **6**: p. 19233.
21. Albright, M.B.N., R. Johansen, D. Lopez, V. Gallegos-Graves, B. Steven, C.R. Kuske, and J. Dunbar, *Short-Term Transcriptional Response of Microbial Communities to Nitrogen Fertilization in a Pine Forest Soil.* Appl Environ Microbiol, 2018. **84**(15).

22. Georganas, E., R. Egan, S. Hofmeyr, E. Goltsman, B. Arndt, A. Tritt, A. Buluc, L. Oliker, and K. Yelick, *Extreme Scale De Novo Metagenome Assembly.* https://arxiv.org/pdf/1809.07014.pdf, 2018.