

81. Expanding the Toolkit for Metagenomics, Implementing in KBase, and Applying It to the Study of the Effects of Experimental Warming in Midwestern and Alaskan Soils

James R. Cole^{1*} (colej@msu.edu), Qiong Wang¹, Jordan Fish¹, Mariah Gilman¹, Konstantinos T. Konstantinidis², Luis M. Rodriguez-R², Liyou Wu³, Zhili He³, Yiqi Luo³, Edward A.G. Schuur⁴, James M. Tiedje¹, Jizhong Zhou³

¹Michigan State University, East Lansing; ²Georgia Institute of Technology, Atlanta; ³University of Oklahoma, Norman; ⁴ Northern Arizona University, Flagstaff

<http://ieg.ou.edu/>

Project goal: The overall goal of this project is to advance system-level predictive understanding of the feedbacks of belowground microbial communities to multiple climate change factors and their impacts on soil carbon (C) cycling processes. The main objectives of this integrative project are to (i) determine the responses of microbial community structure, functions and activities to climate warming, altered precipitation and soil moisture regime in a tundra and temperate grassland ecosystem; (ii) determine temperature sensitivity on recalcitrant C decomposition; (iii) determine the microbiological basis that is underlying temperature sensitivity of recalcitrant C decomposition; and (iv) develop integrated bioinformatics and modeling approaches to scale information across different organizational levels. This work focuses on project utilization and integration with DOE KBase resources.

As part of this project we have developed a suite of tools to help us understand the impact of multiple climate change factors on soil carbon cycling processes. We are in the process of porting these tools to the KBase environment to allow us to better share data and methods between our geographically-separated team, thereby providing rapid access to our data and tools, through KBase, to the broader community of researchers involved in understanding microbial community processes relevant to climate change variables.

In May 2014, two of our developers traveled to Argonne National Labs for an intensive two-day work session to gain hands-on experience with the new KBase deployment procedures and the new central KBase infrastructure components: Shock, a data management system for storing and sharing large data, and AWE, a workflow management system for job scheduling. While at Argonne, we were able to integrate a synchronous RDP Classifier (Wang et al., 2007) developmental version into the new KBase infrastructure. This service accepts input files that were uploaded to the Shock server, executes the RDP Classifier command on the server host and returns results back to users. We have also developed a novel tool called Nonpareil to estimate the sequencing coverage of a metagenomic dataset, i.e., what fraction of the total DNA/species diversity of a microbial community was sampled by a metagenomic dataset, and predict the sequencing effort required to achieve nearly complete community coverage based on the redundancy of shotgun metagenomic reads (Rodriguez-R and Konstantinidis, 2014). This and other tools were developed to overcome several challenges we faced when analyzing soil metagenomes (i.e., the tools were developed to fulfill practical needs). More recently, we have worked with KBase personnel to further increase the computational efficiency of Nonpareil for a large metagenomic database (>100 Gb per dataset) using a kmer approach pioneered by the KBase team, and make Nonpareil available to the scientific community as part of KBase. The release of the tool through KBase is scheduled for later in 2015.

We are proceeding with tool development with “ready for KBase” as a top design priority. We are continuing to implement and improve our new and existing big data analysis tools to be faster and more memory efficient (Cole et al., 2014). We continue to make the source code of our tools available on

GitHub (a KBase deployment requirement), and make them easy to access and install. We continue to produce the detailed documentation and workflow examples required by KBase and that allow the tools to be used more efficiently by the user community. In addition to the tools we previously targeted for KBase integration, we have developed a novel method for assembling specific protein-coding genes of interest. Metagenomics can provide important insights into microbial communities. However, assembling even modest metagenomic datasets with traditional methods has proven to be very computationally challenging. This method uses a combined graph structure and assembles more, longer and better quality gene contigs. This method is implemented as an open source software package called Xander (Wang et al., 2015) and is available at <https://github.com/rdpstaff/RDPTools>.

References

1. Cole, J. R., Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras- Alfaro, C. R. Kuske, and J. M. Tiedje. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 42(1):D633-642; doi: 10.1093/nar/gkt1244 [PMID: 24288368]
2. Rodriguez-R, L. M., K. T. Konstantinidis. 2014. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics* 30(5):629-635.
3. Wang, Q, G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl Environ Microbiol.* 73(16):5261-5267; doi: 10.1128/AEM.00062-07 [PMID: 17586664]
4. Wang, Q., J. A. Fish, M. Gilman, Y. Sun, C. T. Brown, J. M. Tiedje and J. R. Cole. Xander: gene-targeted metagenomic assembler. Submitted.

Funding Statement: This research was supported by the Office of Science (BER), U.S. Department of Energy, Biological Systems Research on the Role of Microbial Communities in Carbon Cycling Program (DE- SC0010715) with additional contribution from the DOE Great Lakes Bioenergy Research Center (BER DE-FC02-07ER64494).