Developing a Toolbox for Soil Metagenomes

Adina Howe, [1]* (howead@msu.edu), Erick Cardenas, Jordan Fish, Jiarong Guo, Aaron Garoutte, Jason Pell,[1] Arend Hintze,[1] James Cole,[1] Patrick Chain,[3,4] Susannah Tringe,[3] Janet Jansson,[2,3] C. Titus Brown,[1] and **James M. Tiedje**[1]

[1]Michigan State University, East Lansing, MI; [2]Lawrence Berkeley National Laboratory, Berkeley, CA; [3]Joint Genome Institute, Walnut Creek, CA; [4]Los Alamos National Laboratory, Los Alamos, NM

http://www.glbrc.org

Project Goals: Our research supports a goal of the Great Lakes Bioenergy Research Center which is to develop sustainable biofuel practices by optimizing soil, microbe, and plant interactions. We have developed tools for soil metagenome analysis which enables us to explore the genetic and population components of soil and rhizosphere communities, with special emphasis on ecological services. Our tools focus on (1) scalable de novo metagenomic assembly, (2) gene-targeted assembly, (3) comparing 16S rRNA gene phylogenetic profiles from gene-targeted and metagenomic sequencing, and (4) detecting the presence of known soil genomes within soil metagenomes.

Metagenomic de novo assembly with digital normalization and Khmer
Assembly of metagenomic datasets is limited to the diversity of the metagenome and available computational memory. We have developed approaches for scaling de novo metagenomic assembly through a data reduction process we call "digital normalization" and a "divide and conquer" strategy of breaking apart large metagenome assembly graphs. We initially reduce the size of the metagenomic dataset by eliminating redundant, high-coverage short-reads within a dataset using a single-pass, constant memory algorithm. This normalization reduces the metagenome dataset size so that it has an increasingly even distribution of read-coverage. Next, we partition disconnected components of the assembly graph using a probabilistic representation data structure called a Bloom filter. These approaches, combined, enable the scalable assembly of very large soil metagenomic datasets. We present the assemblies of two soil metagenomes assembled within less than 66 Gb of memory.

HMMgs: gene-targeted assembly from metagenomic datasets
HMMgs is a tool for assembling protein-coding sequences of targeted genes from large metagenomic datasets. Rather than doing a global de novo assembly, HMMgs enables the assembly of only contigs likely to code for a specific gene of interest. We use Hidden Markov Models (HMMs) to guide the local assembly. Starting assembly points are chosen by identifying specific sequences present in the reads which are similar to reference sequences for the gene of interest. Advantages to this approach are that it requires significantly fewer resources than global assembly and tends to have higher specificity than read-based approaches. We demonstrate and present the successful assembly of *nifH* and *rplB* genes from three Illumina GAII soil metagenomes.

Phylogenetic profile comparisons between pyrosequencing and shotgun sequencing
Pyrosequencing of the SSU rRNA genes amplified directly from environmental samples is now the standard procedure for surveying microbial community structure. In contrast, metagenomic shotgun sequencing does not depend on gene-targeted primers and amplification and thus is not affected by primer bias or chimeras, respectively. Furthermore, shotgun sequences provide opportunities to detect other genes in the ribosomal RNA (rrn) operon, including large subunit genes. Short read lengths and efficiency of annotating enormous number of reads, however, limit the use of metagenomic reads. We compared the phylogenetic profiles of the SSU rRNA genes in pyrotag and shotgun sequencing data from the same soil samples. Pyrosequencing reads were processed using the Ribosomal Database Project Pyrosequencing Pipeline (RDP) and Mothur. SSU rRNA gene fragments in shotgun reads were identified by alignment to known SSU rRNA genes and subsequently classified by the RDP classifier. We found that short reads aligned to the primer regions targeted by our pyrotag data shared high sequence similarity but also detected bases that were not conserved in a few small clades. These sequences were used for evaluating and improving SSU rRNA gene primers for future gene-targeted sequencing efforts. At the phylum-level, community profiles were similar for pyrotag and shotgun data, suggesting that our method for identifying SSU rRNA gene fragments in shotgun data is effective. Our approach also revealed some taxa missed by pyrosequencing primers and supports the use of shotgun metagenomic sequencing for detecting 16S rRNA genes.

Analysis of genomes detected by mapping raw reads to known genomes
To rapidly screen the soil metagenomes to access information about the microbial community and to validate de novo assemblies, we have aligned raw shotgun reads to known reference genomes. A reference genome collection was constructed from 492 complete genomes covering 19 different phyla was retrieved from the NCBI genome database based on metadata in the Gold Database on August 19th, 2011. Microorganisms were manually selected for originating from soil or a similar habitat. Alignment of reads to reference genomes was performed using Bowtie. The coverage of each genome was then calculated for each gene and for the total mapped genome. Out of the 2.7 billion reads from soil growing biofuel crops, 33 million of were successfully mapped to our soil reference genome database (1.28%). Mapped reads ranged from 0.09% to 4.75%, and were more prevalent in rhizosphere samples. The most frequently detected organisms were mostly *Proteobacteria* of which many are known rhizosphere dwellers. Community structure based on detected genomes was similar but not identical to that inferred from 16S rRNA gene pyrosequencing. This is likely due to bias of the databases toward Proteobacteria and human-related organisms. In general average genome coverage was a good predictor of the percentage of mapped reads that were successfully aligned to the contigs.

Sequencing was funded and completed by the Joint Genome Institute, a DOE Office of Science User Facility. Other work was funded by the Great Lakes Bioenergy Research Center (DOE BER Office of Science); the National Institute of Food and Agriculture, Amazon AWS Education; and the National Science Foundation.