Strategies to enable the assembly of soil metagenomes

Adina Howe, [1]* (howead@msu.edu), Jason Pell,[1] Arend Hintze,[1] James M. Tiedje,[1] and **C. Titus Brown[1]**

[1]Michigan State University, East Lansing, MI

http://www.glbrc.org

Project Goals: A goal of the Great Lakes Bioenergy Research Center is to develop sustainable biofuel practices by optimizing soil, microbe, and plant interactions. Soil because of its great diversity has been identified as a "Grand Challenge" for exploring microbial communities through metagenomic sequencing. Our goal has been to develop effective strategies to analyze and assemble very large and complex metagenomic datasets from soil, thus creating opportunities to explore the biological basis and ecological services of soil microbial communities.

The development of next-generation short-read sequencing technologies has allowed us to sequence soil microbial communities to unprecedented depths. We now have extremely large soil metagenomes, which because of their numbers and short read lengths cannot be analyzed with traditional genomic tools. A de novo metagenomic assembly approach offers several solutions. It significantly reduces the data size by collapsing numerous short reads into fewer contigs while providing longer sequences containing multiple genes and operons. Furthermore, it does not rely on the availability of reference genomes. De novo assembly, however, relies on the ability to store the information on the connectivity between all sequences in a dataset within an assembly graph. Thus, soil assembly is challenged by extremely high sequence diversity, uneven sequencing coverage, sequencing errors and biases, and the availability of large computational resources. We have developed novel approaches to enable soil metagenome assemblies through data reduction, scalable assembly graph representations, and removal of sequencing errors and biases.

Our initial step is to reduce the size of the metagenomic dataset by normalizing the average coverage of the soil metagenome using an approach we term "digital normalization." We eliminate 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. Comparing assemblies before and after digital normalization for an *E. coli* genome (50x coverage normalized to 5x), we found that assemblies were > 99% similar despite eliminating 90% of the reads. Similar results were observed for a subset of a soil metagenome.

Another strategy for data reduction is to partition disconnected components of the assembly graph. To do this, we have implemented a probabilistic representation of the assembly graph using Bloom filters. Bloom filters are memory-constant data structures that can be used to store and traverse the assembly graph. As we evaluated the use of Bloom filters to represent large assembly graphs, we found that despite the presence of false positive nodes and edges in the Bloom filter representation, it is effective in dividing its disconnected components.

Using Bloom filters, we were able to partition several metagenomic datasets into millions of disconnected assembly subgraphs. Among these subgraphs, we consistently found a single,

dominant partition consisting of 5 to 76% of metagenomic reads. Characterizing the sequences and connectivity within this dominant partition, we identified position-specific biases within sequenced reads suggesting the presence of spurious connectivity within metagenomes. Using a systematic traversal algorithm, we could identify and remove highly connecting sequences from this partition. We found that the filtering of these sequences not only removes potential sequencing artifacts but also improves assemblies (as demonstrated in simulated datasets) and breaks apart the largest partition allowing for scalable assembly. Applying this partitioning approach to a soil metagenome (30 million reads), we decreased assembly memory requirements by 8-fold.

In conclusion, we have developed approaches that can be applied to the assembly of the growing amounts of soil metagenomic sequencing data. Our approach results in numerous smaller datasets which can be analyzed and/or assembled independently (with separate parameters) and in parallel and subsequently combined into a final assembly for a metagenome.. Furthermore, many of our methods can be extended and applied to other sequence analyses (i.e. transcriptomes).

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.