Methods for high-throughput massively parallel soil ecofunctional gene analysis

James R. Cole[1]* (colej@msu.edu), Jiarong Guo,[1] Leo Tift,[1] James M. Tiedje,[1] and **Phil Robertson**[1]

[1]Michigan State University, East Lansing, Michigan 48824

Project Goals:
To improve the sustainable development of bioenergy, our project characterizes beneficial microbes in three important bioenergy crop systems and their impacts on critical biogeochemical processes, especially the nitrogen cycle, and ultimately explores ways to manage such microbiomes for improved biofuel sustainability.

http://www.glbrc.org

In analysis of biofuel crop soil microbiomes we are interested in targeting both genes providing a taxonomic/phylogenetic analysis of soil community microbial members and an analysis of ecofunctional genes involved in microbe-plant interactions, including genes coding for ecological processes important to plant and soil health. We are targeting these genes through both primer-targeted (PCR-based) and untargeted (shotgun) sequencing methods and analysis.

Standard taxonomic analysis has relied on amplification and sequencing of SSU rRNA genes. This method is well developed and inexpensive but has several drawbacks. First, the SSU rRNA genes are very slowly evolving and have limited resolving power for taxa below the level of genus; second, the SSU gene is often present in multiple copies per cell, presenting difficulties in calculating relative abundance of different organisms. Because there are fewer evolutionary constraints, many protein-coding core genes are more rapidly evolving than the SSU rRNA gene yet share many of the same advantages as a taxonomic marker, e.g., are present in all organisms and are unlikely to be horizontally transferred. In addition, most are normally present in a single copy. For our studies we have chosen the *rplB* gene coding for ribosomal protein L2. Examining completed bacterial genome sequences, we found that, as expected, the *rplB* gene shows more change than the SSU rRNA gene so is better able to resolve at the lower ranks, which are more ecologically relevant (Fig 1).

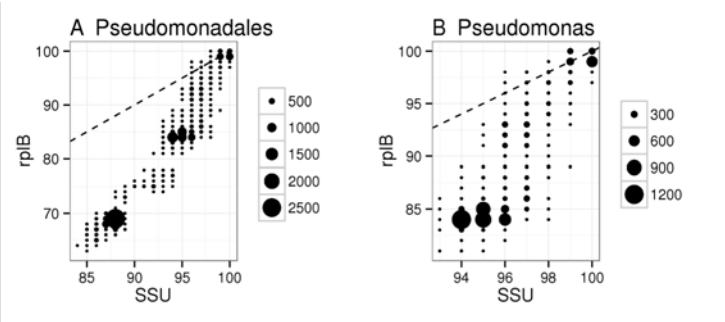

Figure 1. A. All completed genomes in Pseudomonadales are included in pairwise comparison; **B.** All completed genomes in Pseudomonas are included. The dashed line is y = x. Size of dots indicates number of genome pairs that share the same SSU rRNA gene identity and *rplB* identity.

However, it can be difficult to use these protein-coding core genes as markers because the known PCR primers have phylogenetic biases. To circumvent this problem, we have used the Xander

tool previously developed for this project (Wang et al., 2015. Xander: employing a novel method for efficient gene-targeted metagenomic assembly. *Microbiome* 3:32) to assemble and annotate core phylogenetic marker genes beyond the SSU rRNA gene. This Xander tool uses a Combined Weighted Assembly Graph (CWAG) that combines nodes from a standard De Bruijn graph representation of shotgun data with the states (nodes) of a Hidden Markov Model (HMM) for the gene(s) of interest. The HMM information adds weights to the graph edges and allows us to search for the gene(s) of interest using standard graph-theory path-finding algorithms on the CWAG. We used Xander to assemble *rplB* genes from 21 soil metagenomic data sets, seven replicates from each of three biofuel crops. We found that the *rplB* gene provided much better separation of the three biofuel crops than SSU rRNA genes isolated from the same metagenomic datasets (Fig 2).

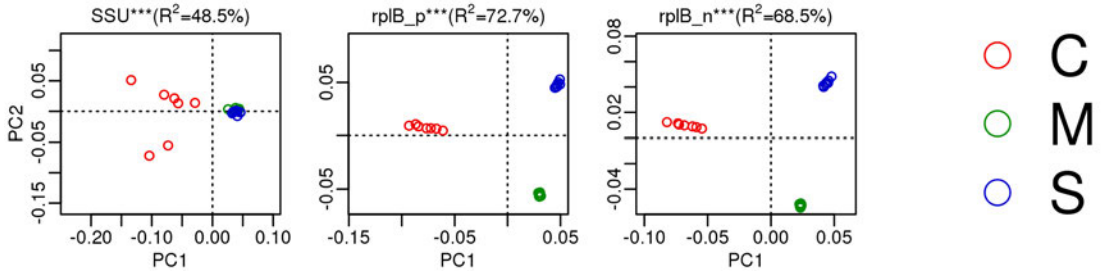

Figure 2. Comparison of SSU rRNA gene and *rplB* in beta diversity analysis (ordination) using large soil metagenomes. Both genes show microbial community in corn (C) rhizosphere is significantly different from those in *Miscanthus* (M) and switchgrass (S). Additionally, *rplB* (both nucleotide and protein) separate communities of *Miscanthus* and switchgrass..

Many important ecofunctional genes are relatively rare and present in only a small fraction of bacterial cells. Bulk metagenomics either fail to accurately quantify changes in these genes or require excessive sequencing depth at greatly increased cost for sequencing and data processing. Protein-coding genes are, in general, less conserved than structural RNA genes, meaning that often no single probe or primer pair is able to target a gene's full range of diversity. We have been developing and testing an efficient high-throughput primer design tool to develop multiple PCR primer sets for each targeted gene. Our tool helps with the design of multiple primers from potentially large reference sets of 3000 sequences or greater. We cast the problem as a variant of the well-known "maximum coverage problem" from computer science. Since this problem has no practical exact solution, we use a "greedy" algorithm to choose a set of primer pairs from the candidates that maximizes the diversity covered by the primer sets. It requires an aligned set of reference sequences as input, with an optional phylogenetic tree for diversity weighting, an optional amplicon size range, and a maximum number of primer pairs to be developed per gene.

During tool testing, we developed new primer sets for nitrogen cycling genes (*amoA*, *nifH*), recalcitrant carbon degradation genes (*cutC*, *cntN*), antibiotic resistance genes (*tet_sul2*, *tetA-G*), an integrase gene involved in mobile elements (*intI1*) and a microbial gene involved in reducing plant stress (*acdS*). We have experimentally validated a set of three non-degenerate primer pairs targeting *cntN* sequences. The sequencing results showed adequate sensitivity, satisfactory amplicon size, and 99% PCR efficiency with the three primers used. This tool is also being employed in the design of primers targeting ACC deaminase for investigations into the role of disease suppressive soil microbial communities in promoting plant health. The current tool is already in use by several research groups. We have developed the tool to be "KBase ready" and intend to help integrate this functionality both into our FunGene and into KBase.