MicroCOSM: Phylogenetic Classification of Metagenomic Data Using Microbial Clade-oriented Sequence Markers

Dylan Chivian[1,2]* (DCChivian@lbl.gov), Paramvir S. Dehal[1,2], **Adam P. Arkin**[1,2,3]

[1]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov ; [2]Lawrence Berkeley National Laboratory, Berkeley, CA; [3]University of California, Berkeley, CA

The VIMSS/ESPP2 project requires understanding of the microbial communities at contaminated field sites and, among other methods, will employ metagenomics in this endeavor. Metagenomics projects that seek to elucidate the population structure of microbial ecosystems are faced with the related computational challenges of classifying the sequences obtained and quantifying which organisms are present within a sample. Individually low-proportion species usually make up a large fraction of microbial communities, complicating their classification and quantification using traditional phylogenetic marker approaches. Such species usually don't yield sufficient read depth to assemble into longer sequences, leaving fragments that rarely contain traditional markers such as the small subunit (SSU) rRNA gene. BLAST-based approaches for analysis of metagenomic sequences [1] compensate for this rarity of traditional markers, but may be confounded by genes that are subject to horizontal transfer or duplication. Another approach instead makes use only of reliable non-transferred single-copy genes [2] to classify and quantify the organisms present within a sample, but the application has so far been limited to the use of a fairly small set of universal genes found in all organisms. In this work, we have extended the latter approach, boosting the set of reliable marker genes from only about 30-40 universal genes to several hundred by identifying sets of single-copy genes that are not subject to inter-clade horizontal transfer through investigation of finished bacterial and archaeal genomes. These clade-oriented sequence markers allow for a method, which we have named "MicroCOSM", that greatly increases the probability that a marker will be found in any given sequence and therefore offers improved coverage for phylogenetic classification and quantification of microbial types in an environmental sample.

[1] Huson D.H., Auch A.F., Qi J., Schuster S.C. (2007) "MEGAN analysis of metagenomic data." *Genome Res*. **17**(3):377-86.

[2] von Mering C., Hugenholtz P., Raes J., Tringe S.G., Doerks T., Jensen L.J., Ward N., Bork P. (2007) "Quantitative phylogenetic assessment of microbial communities in diverse environments." *Science* **315**(5815):1126-30.