A Method for Circularizing Microbial Genomes from Metagenomics Data

L.M. Lui (lmlui@lbl.gov)[1*], T. Nielsen[1], H.J. Smith[2], F. von Netzer[3], E.L-W. Majumder[4], J.V. Kuehl[1], F. Song[1], A. Sczesnak[1,5], M.P. Thorgesen[6], X. Ge[6], F.L. Poole[6], C.J. Paradis[7], K.F. Walker[8], K.A. Lowe[9], D.C. Joyner[9], D. Ning[9], M. Rodriquez, Jr.[8], A.B. Aaring[1], B.A. Adams[8], D. Williams[8], J.D. Van Nostrand[10], G.M. Zane[11], M.W.W. Adams[6], J. Zhou[10], R. Chakraborty[1], J.D. Wall[11], D.A. Stahl[3], T.C. Hazen[8,9], M.W. Fields[2], AP Arkin[1,5], **PD Adams[1]**

[1]Lawrence Berkeley National Lab, Berkeley CA; [2]Montana State University, Bozeman MT; [3]University of Washington, Seattle WA; [4]State University of New York, Environmental Science and Forestry; [5]University of California, Berkeley CA; [6]University of Georgia, Athens GA; [7]University of Wisconsin, Milwaukee WI; [8]University of Tennessee, Knoxville TN; [9]Oak Ridge National Lab, Oak Ridge TN; [10]University of Oklahoma, Norman OK; [11]University of Missouri, Columbia MO

http://enigma.lbl.gov

Project Goals: ENIGMA -Ecosystems and Networks Integrated with Genes and Molecular Assemblies use a systems biology approach to understand the interaction between microbial communities and the ecosystems that they inhabit. To link genetic, ecological, and environmental factors to the structure and function of microbial communities, ENIGMA integrates and develops laboratory, field, and computational methods.

Metagenomics facilitates the study of the genetic information from unculturable microbes and complex microbial communities, but achieving complete microbial genomes (*i.e.*, circular) from metagenomics data is difficult because of inherent qualities of short read sequencing data and assembly and binning methods currently available. To our knowledge, only 62 circularized genomes have been assembled from metagenomics data despite the thousands of datasets that are available. We believe that circularized genomes are important for (1) building a reference collection as scaffolds for future assemblies, (2) providing complete gene content of a genome, (3) confirming little or no contamination of a genome, (4) studying the genomic context of genes, and (5) linking protein coding genes to ribosomal RNA genes to aid metabolic inference in 16S rRNA gene sequencing studies. We developed a method to achieve circularized genomes using iterative assembly, binning, and read mapping. In addition, this method exposes potential misassemblies from k-mer based assemblies. We chose species of the Candidate Phyla Radiation (CPR) to focus our initial efforts because they have small genomes and are only known to have one copy of the ribosomal RNA genes. From our work, we present 42 CPR genomes and one Margulisbacteria genome from 19 published datasets and from ENIGMA sequencing of sediment and groundwater samples from Oak Ridge National Lab Field Research Center. We demonstrate findings that would likely be difficult without circularized genomes, including that ribosomal genes are likely not operonic in the majority of CPR, diverged forms of RNase P in CPR, and presence of megaplasmids in the datasets.

This material by ENIGMA- Ecosystems and Networks Integrated with Genes and Molecular Assemblies a Scientific Focus Area Program at Lawrence Berkeley National Laboratory is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological & Environmental Research under contract number DE-AC02-05CH11231