ENIGMA Long Read Sequencing and Assembly for Microbial Genomes: Current Capabilities and Future for Metagenomics and KBase Integration for Assembly

Lauren M. Lui*[1] (lmlui@lbl.gov), Torben N. Nielsen*[1] (torben@lbl.gov), John-Marc Chandonia[1], Adam P. Arkin[1,2], **Paul D. Adams[1]**

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]University of California, Berkeley, CA

https://enigma.lbl.gov

Project Goals: ENIGMA (Ecosystems and Networks Integrated with Genes and Molecular Assemblies) uses 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.

Achieving a causal understanding of a microbial system requires mapping mechanisms by which organisms grow, cooperate, and compete in complex environments. These mechanisms include ecological phenomena and abiotic factors that influence behavior and survival. One of the critical requirements for reaching this level of understanding is fully resolving the genomes of the community, including plasmids and viruses, so that the functional roles specified by their genomes can be assayed and discovered. While the challenges of gene functional annotation and linking genotype and phenotype loom beyond simply obtaining genomes, the underlying challenge at the present remains to generate high-quality genomes for both isolates and metagenomes. The base genome along with its relative abundance constitute the most important foundational data needed to infer and parameterize models of microbial system dynamics.

Despite advances in sequencing technology throughput, until recently it has been very difficult or impossible to completely finish genomes from both isolates and microbial communities. Finishing genomes is difficult and laborious without the use of reads that are longer than any repetitive DNA element, which can be thousands of base pairs. In addition, for metagenome-assembled genomes, it is challenging to associate plasmids with their host chromosomes. Both these barriers can be overcome using new technologies in long read sequencing from Pacific Biosciences and Oxford Nanopore Technologies. These technologies can produce read lengths in the multiple thousands and can resolve DNA methylation modification. Error rates in calling nucleotide identity have been dropping as well, reducing the need for more accurate (Illumina) short reads for polishing. Long reads enable effective assembly of replicons and can even permit strain-level resolution variants. Methylation profiling can often tie multiple replicons such as genomes and plasmids together in the same organism even when assembled from a complex mixture of organisms. Full genome assemblies wherein all replicons are associated accelerates studies of phylogeny, adaptive evolution, and facilitates better assessment of base genetic capability. While this technology is progressing it still remains challenging to apply it effectively to more complex samples of natural communities and diverse enrichments.

The ENIGMA SFA has spent time developing pipelines for sequencing and assembly of long read data from microbial isolates and metagenomes to help achieve the goal of casual microbial ecology. We have developed the capability to isolate diverse organisms, to extract the high molecular weight (HMW) DNA needed for single molecule long read sequencing, and to perform the sequencing using Oxford Nanopore Technologies MinION sequencers. For example, we have successfully made long read libraries using DNA extracted from ENIGMA groundwater and sediment samples. From one sample we recovered more than 30 finished grade genomes, including two 6 Mb genomes, as well as what we believe to be a bacteriophage larger than 1 Mb, which would make it the largest bacteriophage genome known (currently the largest identified phage is ~735kb). This is from one sample; from nearly all short read metagenome studies there are only approximately 100-200 fully complete microbial genomes.

We developed computational pipelines to process long read sequencing data since ENIGMA is generating next-generation sequencing datasets at scale. To characterize the microbial diversity and activity at the Oak Ridge Reservation at Oak Ridge National Laboratory, ENIGMA anticipates isolating thousands of bacteria and archaea, as well as generating spatio-temporal series of fully resolved enrichments and metagenomes from the site. These sequencing projects assist the goals of linking genotype to phenotype and understanding the temporal, dynamic, and complex factors influencing microbial community structure and activity at our site. ENIGMA uses isolates to help link genotype to phenotype by analyzing genomes in conjunction with transposon mutant libraries, metabolomics, and growth condition data. High quality genomes and metagenomes are essential for these types of experiments and ENIGMA science.

We are currently adding new functionality to DOE Systems Biology KnowledgeBase (KBase) by implementing tools for using long read data for assembly of isolates, assembly of metagenomes, and methylation detection. We are developing workflows within KBase to make these tools more broadly available across the ENIGMA SFA and to other scientists, especially for scientists that do not specialize in computational methods. These apps and workflows will enable ENIGMA, as well as other DOE SFAs and microbiologists to (1) address scientific questions that would otherwise be infeasible with isolate and metagenome assemblies using only short reads, (2) track provenance of data and methods used for assembly, and (3) share assemblies across the SFA for collaborations. By providing this new functionality in KBase, we will also provide a foundation for further extensions in KBase to support developments in long read technology.

This material by ENIGMA- Ecosystems and Networks Integrated with Genes and Molecular Assemblies a Science 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