62. Rapid Binning and Metabolic Profiling of Subsurface Microbial Community Metagenomic Data via an Interactive Online Knowledgebase

Brian C. Thomas, Andrea Singh, Laura A. Hug* (laura.hug@berkeley.edu), and **Jillian Banfield**

University of California - Berkeley, Berkeley CA

Project goals:
Our goal is to develop an online resource, ggKbase, for the management and rapid analysis of microbial community data, including metagenomic and transcriptomic sequences and proteomic profiles. The platform offers a custom system for genome binning, metabolic pathway curation, and community composition analysis. Importantly, these functions are combined with visualization tools that help guide analysis and interpretation of complex datasets, as well as visually summarize the results for publication. The ggKbase platform is useful for researchers interested in the rapid binning and metabolic assessment of metagenomic data.

URL: http://ggkbase.berkeley.edu/

Abstract:
The scientific focus of our research is to develop a predictive understanding of the microbiology of the subsurface, including the roles of microorganisms in carbon cycling. The questions metagenomic datasets can address are multiple: what organisms are present in a community, what is their relative abundance, and how do these change over time and space or in relation to changing physiological conditions. Beyond taxonomic profiling, metagenomics allows prediction of the metabolic potential of the community, including which processes may be occurring and through which intermediates. Transcriptomics and proteomics can be used to confirm the active metabolic processes and identify the key members within a community for nutrient cycling. Datasets are growing in size in line with increases in sequencing capacity, and the communities being examined are likewise scaling in complexity and diversity. Thus, there are major challenges relating to efficiency and accuracy of data analysis.

The ggKbase platform is designed as an interactive, online environment for the simultaneous and partially automated analysis of hundreds of genomes and associated omic data. ggKbase offers a suite of tools for rapid assignment of assembled fragments to organisms (binning) and metabolic prediction. Intrinsic to these analysis options are interactive visualizations that allow fast and intuitive examination of the data, streamlining the analysis process and providing summaries that can be used in publications. We will present several examples of applications of ggKbase for binning (Fig. 1) and metabolic analysis (Fig. 2) of complex subsurface communities, containing hundreds to thousands of organisms, many of which are highly novel compared to existing genome databases.

To date, the ggKBase platform has been extensively used for metagenomic analyses on complex communities from aquifer sediment, acid mine drainage, and the human gut. These projects have allowed tool development informed by user requirements from manual investigations of the data. The ggKbase is now being transitioned to support community users on a wide scale.

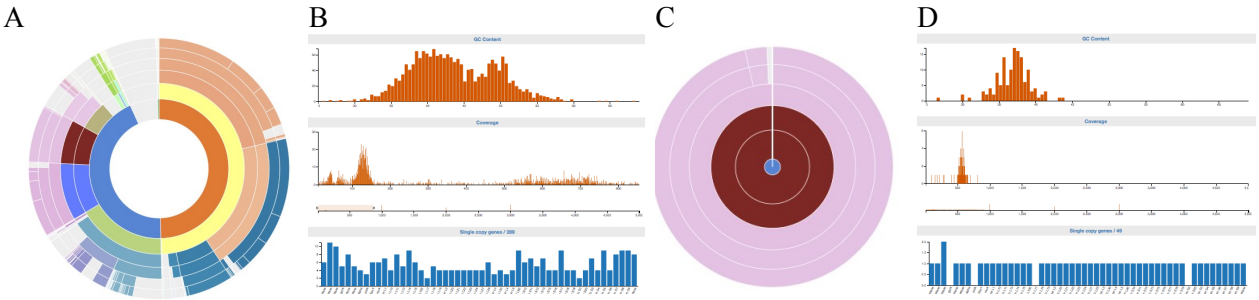

Figure 1: A) Interactive display of the phylogenetic profile of unbinned metagenomic data. Concentric rings relate to the taxonomic level to which a contig can be assigned based on the phylogenetic profile of encoded genes. B) Histograms profile the GC content, coverage, and single copy gene content of the whole dataset. Phylogenetic groups can be chosen, and subsets of these groups assigned to bins using interactive controls that select specific GC content and/or coverage ranges. For example, clicking a region in the taxonomic wheel in A (C) shows that a bin with consistent GC and coverage profile has been generated, and the single copy gene profile indicates that the putative bin could contain a relatively complete genome (D).

Figure 2: Example of the use of dataset-wide lists to profile both metabolic potential and genome completeness in binned data