Abstract: Diversity of functional genes in deeply branching uncultured microbes, Y1

The microbial world primarily consists of taxa that have never been grown in culture (Steen et al., 2019; Lloyd et al., 2018). These taxa are a potential wellspring of undiscovered genomic function, but the quantity of novel functional genes in an organism does not necessarily track the phylogenetic distance of that organism from cultured lineages. In order to test whether phylogenetically divergent microbes also contain more novel genetic material, we measured the sequence similarity between each gene in approximately 15,000 genomes of bacterial and archaeal isolates, metagenome-assembled genomes, and single-cell genomes, to the most similar gene in the SwissProt database of genes of well-known function. We compared these sequence similarities between predicted proteins to the phylogenetic distance between each organism and its closest cultured relative, allowing us to assess whether uncultured phyla contain more genetic novelty than cultured phyla. A similar analysis allowed us to measure the quantity of genetic novelty within ecosystems. Finally, we measured the typical distance between genes of unknown function within taxa and within environments in order to test whether some phyla or environments contain more, different kinds of genes than others. We found that, in general, uncultured phyla contain greater genetic novelty than cultured phyla, suggesting that uncharacterized organisms may be a source for novel functions for determining ecosystem services or biotechnological applications.

Stage 1 work was performed using HPC-BLAST (Sawyer et al., 2015) to compute similarity scores of the multiple sets of gathered metagenomes against the SwissProt (Boutet et al., 2007; Bateman, 2019) database of genes. HPC-BLAST is an implementation of the BLASTP (Gish and States, 1993) algorithm which allows for efficient deployment of blastp on a modern hybrid distributed-multicore shared memory architecture. Due to the size of the comparison dataset, the code is most efficient when run in ensemble mode, balancing memory requirements with the number of cores on each node.

Stage 2 involves the aggregation and analysis of the results of the BLAST results from stage 1.

Stage 3 involves the comparison of all metagenomes to all other metagenomes in our collection. Similar to stage 1, this work requires HPC-BLAST comparisons, but instead of a single comparison database (SwissProt 276MB, 560K sequences), we will compare will all other metagenomes in our collection. So, while the metagenomes vary in size, we can estimate an O(N2) computational cost, where N is the total number of metagenomes in our collection.