Novel Computational Approaches to Investigate Microbial Diversity

By Qingpeng Zhang

Species diversity is an important measurement of ecological communities. Scientists believe that there is a strong relationship between species diversity and ecosystem processes. With novel applications of data structures and the development of novel algorithms, we integrate efficient k-mer counting and read coverage analysis based on digital normalization to create a novel statistical framework allowing for scalable diversity analysis of large, complex metagenomes without the need for assembly or reference sequences. This method is evaluated on multiple large metagenomes from a variety of environments, such as seawater, human microbiome, soil. Given the velocity in growth of sequencing data, this method is promising for analyzing highly diverse samples with relatively low computational requirements. Further, as the method does not depend on reference genomes, it also provides opportunities to tackle the large amounts of unknown “dark matter” we find in metagenomic datasets.