Species diversity is an important measurement of ecological communities. Scientists believe that there is relationship between species diversity and ecosystem processes. In almost all the metagenomics projects, diversity analysis plays an important role to supply information about the richness of species, the species abundance distribution in a sample or the similarity and difference between different samples, all of which are crucial to draw insightful and reliable conclusion. Since we have limited sequencing power and other financial strain, the resulting metagenomics data set from high diversity sample like soil only corresponds to a tiny fraction of the actual genomic content in the sample. The large size of data set and the low coverage make the assessment of microbial diversity of high diversity sample even harder. With the novel applications of data structures and the development of novel algorithms, my research provides the necessary and highly desired computational methods to enable scalable microbial diversity analysis of the complex metagenomes, with further potential to facilitate other analysis like assembly, annotation.

In the past several years, I have helped starting and participated in an effort to develop the khmer software package (`https://github.com/ged-lab/khmer`), for fast and memory efficient online counting of k-mers in sequencing data sets. Unlike previous methods based on data structures such as hash tables, suffix arrays, and trie structures, khmer relies entirely on a simple probabilistic data structure, a Count-Min Sketch. The Count-Min Sketch permits online updating and retrieval of k-mer counts in memory which is necessary to support online k-mer analysis algorithms. On sparse data sets this data structure is considerably more memory efficient than any exact data structure. We conducted extensive analysis on the performance of the counting algorithm and benchmark to compare the performance of the khmer to other k-mer counting packages. We also put a lot of effort into making the khmer paper reproducible with an automatic pipeline (`https://github.com/ged-lab/2013-khmer-counting`). The initial motivation of developing khmer was to count the k-mers in metagenomes for diversity analysis. Now khmer has been widely used for many other purposes, from enabling large scale de novo metagenome assembly to sequencing error detection and correction.

Based on the efficient k-mer counting package khmer, I helped developing digital normalization, which is a single-pass computational algorithm that systematizes coverage in shotgun sequencing data sets, thereby decreasing sampling variation, discarding redundant data, and removing the majority of errors. Digital normalization can substantially reduce the size of shotgun data sets and decrease the memory and time requirements for de novo sequence assembly, especially for large complex metagenome samples, all without significantly impacting content of the generated contigs. The algorithm of digital normalization has been used by many research groups to facilitate their analysis and has been implemented in different tools like Trinity and Illumina's TruSeq pipeline.

Furthermore, I integrate efficient k-mer counting and a novel de Bruijn graph mapping method based on digital normalization to develop a new method to allow for scalable diversity analysis of large, complex metagenomes. A novel concept - IGS (informative genomic segment) is proposed to represent the unique information in a metagenomics data set. The IGSs can be used as a replacement of OTUs to be the cornerstone for diversity analysis of whole shotgun metagenomics data sets. The abundance of IGSs in different samples can be retrieved by mapping the reads to de Bruijn graph database. I have evaluated this method on multiple metagenomes from a variety of environments (e.g., human gut, human body part, soil in collaboration with James Tiedje, ballast water viromes in collaboration with Joan Rose). Given the velocity in growth of sequencing data, I believe that this method is promising to highly diverse samples with relatively low computational requirements. Further, as they do not depend on reference genomes, these methods also provide opportunities to tackle the large amounts of unknown ”dark matter” we find in metagenomic datasets.

Generally I am strongly interested in developing and applying computational methods to guide large scale efforts of using sequencing technologies as a tool to answer biological questions. I am especially thrilled by the computational challenge presented in the field of metagenomics. For all the methods I have been working on, like k-mer counting, de Bruijn graph mapping, IGS diversity analysis, I have being working really hard to make them more efficient, more powerful, more scalable, which has facilitated the metagenomic research in different ways. But there is still a long way to go while facing much bigger metagenomic data sets and more complex metagenomic samples. Going forward, I plan to integrate those methods I have been working on, with other powerful machine learning methods, other efficient data structures or algorithms to tackle more problems, like binning approaches, functional annotation, and phylogenetic analysis, trying to make the approaches more efficient , more powerful and more scalable, as well. These problems are excellent examples to represent the computational challenges in biological research. Working on these problems can help microbial ecologist as well as all humanity to acquire more knowledge about the microbial world, it can also demonstrate the power of efficient computational approaches to facilitate actual scientific research.