Microbes are everywhere, living in complex communities with enormous metabolic versatility. Microbes modulate and maintain our atmosphere, keep us healthy, and help us grow our food. Additionally, microbes from the natural environment have proven indispensible in our built environments, providing important functions such as wastewater treatment and bioremediation. Although microbial interactions with each other and their surrounding environment provide services essential to our survival, we know very little about the rules of their function. As our society seeks to develop long-term, sustainable management plans for our global resources, it is critically important to understand the services provided by microbes in the environment and how they respond to change. My research provides access to the genetic basis underlying complex interactions in natural and engineered communities, and I intend to elucidate the most relevant microbial drivers of ecosystem health (e.g., stability and resilience). My work is unique and takes advantage of high throughput next generation sequencing, a relatively new technology that has revolutionized the way we investigate the microbial world. Through sequencing, we now have unprecedented access to *all* active genes in microbial communities (rather than a limited number of representative isolated genomes). As one of the few researchers developing tools for making sense of high throughput sequencing data, my work is on the forefront of understanding microbial community dynamics.

My research develops laboratory and computational methods that link local processes to large-scale spatial dynamics to improve our understanding of complicated systems. Specifically, I intend to address the following questions in both natural and engineered environments: (1) *What are the most relevant feedbacks and interactions of microbial processes in Earth's ecosystems and the atmosphere, and how does the nature of these interactions respond to a changing climate?* (2) *What are the relative contributions of microbial genes, individuals, and groups to resulting population function and dynamics?* (3) *How do we combine new technologies with multiple scales of investigation (e.g., from laboratory and field to earth observations) to make robust ecological inferences?* To answer these questions, my research closely couples innovative biological, geochemical, and computational approaches to provide an integrated view of microbial roles in system function. Below, I describe specific research projects exemplifying how I aim to contribute to answering each of these questions.

(1) System services: What are the microbial drivers of carbon cycling in bioenergy soils?

Soils are a large and important reservoir for carbon – storing about twice as much carbon as the atmosphere. Microbial communities play important roles for carbon cycling in the soil as they both catalyze soil carbon mineralization and release CO_2 into the atmosphere. Understanding the microbial drivers of carbon cycling in soils is important, both for managing carbon stocks and for projecting future interactions between the climate and microbial carbon cycling. Research in this area has historically been challenged by the physical and microbial complexity of the soil. To be effective, investigations must address the highly heterogeneous physical structure of soil itself as well as the diversity of the microbial communities throughout. In a mere gram of soil, there are billions of *unidentified* bacterial species, which vary even among soil samples from the same plot. To overcome this obstacle, my research leverages "experimentally partitioned" soils comprised of sieved soil aggregates of varying sizes. Soil is comprised of naturally occurring particles which bind to each other, and these soil aggregates provide a range of pore sizes which are characterized by different abilities to exchange nutrients, air, and water. The experimental separation of the physical soil structure into its constituent aggregates not only reduces the complexity of the soil system but also provides a scale that is consistent with microbiology and tractable to target microbial processes. Genomic sequencing and novel computational approaches have enabled direct access to genes in soil communities. By applying these "omic-based" approaches to soil aggregates, I identified the carbon metabolic profile (the suite of potential carbon cycling genes) of microbial communities in various soil aggregate sizes. Integrating laboratory measurements of enzyme potential (e.g., cellulose

degradation) with molecular estimations of gene abundance, I also observed strong associations between specific genes and enzymes that varied among aggregates of different sizes. Furthermore, I identified specific microbial populations which differentially contribute to various mechanisms of carbon degradation. As part of a multi-year DOE grant, I will lead the next stage of this study, leveraging this existing metagenomic data to exploit metatranscriptomics (identification and quantification of expressed genes) of field soils and engineered model systems (e.g., carbon-enriched soils) to further *quantify* the response of soil communities to varying carbon pools and warming temperatures.

(2) Population dynamics: What is the diversity and role of pathogenic viruses in raw sewage and wastewater treatment effluent?

Viruses are estimated to be the most abundant and diverse type of biological entity in almost every ecosystem and consequently can have a pronounced effect on system function. For example, bacteriophages, the viruses that infect bacteria, play an important role in shaping microbial community structure by causing irreversible population shifts and evolutionary redirection (e.g., through horizontal gene transfer). Furthermore, viruses cause a wide range of disease in humans, animals, and plants. One potential but understudied pathway for the dispersal of viruses is wastewater. Pathogenic viruses are shed in extremely high numbers in fecal matter and discharged into sewage. Yet, we know very little about virus composition in sewage and the effectiveness of wastewater treatment in removing these pathogens. To address this critical gap, I have characterized the diversity of genes from viral DNA extracted from raw sewage, resulting in the identification of 43,890 genes. The majority of these genes originate from known bacteriophages, suggesting dynamic interactions with bacterial hosts with the potential to compromise wastewater microbial communities. Eukaryotic viruses were also identified and originated from animal and plant viruses. Human viral pathogens (including human adenoviruses B and F and polyomavirus JC virus) comprised only a fraction of viral diversity but were present in high abundance in sewage. Moving forward, my research in this domain will include the development of a more extensive gene catalog of the viral community in both raw sewage and wastewater effluent. This reference is critically needed for the study of viruses as the existing database is extremely limited and biased, containing approximately 1100 sequenced phages, 85% of which originate from only 3 bacterial hosts. Furthermore, to better understand host-pathogen interactions, I will also include investigations of the complementing bacterial community with viral targets. The results of this research are an important first step towards identifying viral biomarkers for water quality and management.

(3) Novel methods: How can we access biodiversity in complex systems with available technologies?

Our ability to characterize complex communities is limited by two major factors: how accurately we can sample a diverse community and how completely we can characterize this sample. While technological advances have greatly expanded or ability to sample communities, they have simultaneously ushered in new challenges pertaining to our inability to effectively analyze the results, particularly in complex environments. Research based on sequencing approaches is further hindered by a lack of available genomic references. Application of omic-based approaches to complex environments offers the potential to obtain deep sampling of its genomic content but at the cost of extremely high volumes of data (often greater than 1 Tb for a single sample) requiring considerable computational resources (> 250 Gb of RAM) and specialized tools to analyze. Analysis of such omic-datasets without available (or accurate) references requires the *de novo* assembly of the raw data into its original genomes (or fragments of these genomes). Analogous to building a jigsaw puzzle, *de novo* assembly involves identifying the connections between short genomic fragments (akin to puzzle pieces) in order to create a picture of the microbial community (without the aid of a picture on the front of the puzzle box). This objective is made more daunting because the data typically reflects under-sampled genomic diversity (missing puzzle pieces) as well as sequencing errors (pieces that don't belong to this puzzle). Through my research, I have developed

and successfully implemented methods suitable for assembling *previously intractable* large, complex datasets allowing for adequate characterization of the results of extremely large sequencing projects. In small soil metagenomes (containing 3000 Mbp of sequencing), my approaches reduce the computational requirements for assembly by more than 50-fold; for example, the required memory is reduced from 50 GB of RAM to <2 GB, enabling data analysis using nothing beyond a common personal laptop. I continue to refine these approaches and apply them to build critical assembled gene reference catalogs for exploring the functional potential of multiple ecosystems. Furthermore, I am participating in a funded DOE Knowledgebase collaboration to further develop these approaches (with an emphasis on facilitating metatranscriptomic assembly) for the broader scientific community. I am also working on novel computational approaches to integrate varied sets of –omic sequencing and biogeochemical datasets.

Understanding the mechanisms that control microbial interactions in highly complex communities and the associated implications for system functioning remains a major challenge in the field of environmental engineering. In order to address the aforementioned critical ecological questions, my research aims to provide insight into key systems by integrating multiple contemporary methodologies to characterize the interactions between microbial communities and their physical, chemical, and biological environments. Thus far, I have combined experimental and computational approaches to tackle important yet daunting natural and built environments including soil and wastewater communities. Notably, though, these approaches can easily be extended beyond these specific ecosystems. For example, my investigations into the most relevant microbial drivers of carbon degradation in the soil can identify possible targets for microbial degraders of cellulosic biomass. My experience in engineering, microbiology, ecology, and bioinformatics provide a *unique* and *robust, interdisciplinary framework* for my future research, which will ultimately help us sustainably manage our most critical microbial systems.