Bioinformatics

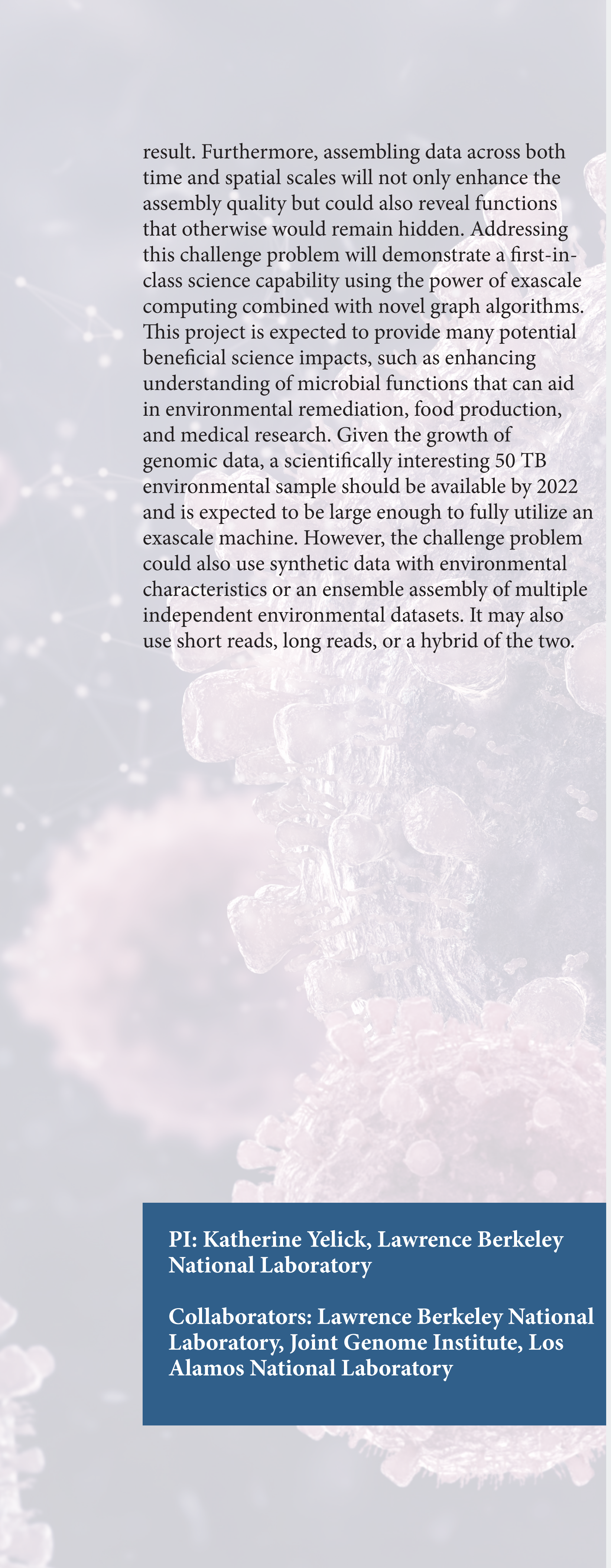
ExaBiome: Exascale Solutions for Microbiome Analysis

Genome sequencing on DNA extracted from microbiomes is used to study the diversity, integration, and dynamics of organisms in the microbiomes. Due to the size and complexity of the datasets involved, assembly and comparative analysis are the most computationally demanding aspect of this branch of bioinformatics. Furthermore, as more data become available, this cost will only grow. The ExaBiome project is developing scalable data assembly and analysis tools to address current needs and, through the use of exascale computing power, provide solutions for anticipated increases in biological data.

Metagenomics—the application of high-throughput genome sequencing technologies to DNA extracted from microbiomes—is a powerful and general method for studying microbial diversity, integration, and dynamics. Since the introduction of metagenomics over a decade ago, it has become an essential and routine tool. Assembly and comparative analyses of metagenomic datasets are among the most computationally demanding tasks in bioinformatics. The scale and rate of growth of these datasets will require exascale resources to process (i.e., assemble) and interpret through annotation and comparative analysis. The ExaBiome project aims to provide scalable tools for three core computational problems in metagenomics: (1) metagenome assembly, which takes raw sequence data and produces long genome sequences for each species; (2) protein clustering, which finds families of closely related proteins; and (3) signature-based approaches to enable scalable and efficient comparative metagenome analysis, which may show, for example, variability of an environmental community over time.

The ExaBiome team has developed a scalable metagenome assembler, MetaHipMer, which scales well on thousands of compute nodes on today's petascale architectures and has already assembled large environmental datasets that had not been possible with previous tools. They continue to work on further scalability improvements across nodes and new node-level optimizations to take advantage of fine-grained on-node parallelism and memory structures including GPUs. MetaHipMer exhibits competitive quality with other assemblers, and the team continues to add innovations and parameters to control various aspects of how data are analyzed, driven by the experience of science teams. MetaHipMer is designed for short reads (Illumina) data, but a second assembler for long reads is also under development and shows even higher computational intensity, which may be a good fit for exascale systems. A second ExaBiome code, HipMCL, provides scalable protein clustering. HipMCL runs on thousands of nodes and has already been used to provide insight on the structure of protein families across hundreds of millions of proteins, a dataset that was previously intractable. These codes and comparative analysis tools use some common computational patterns, including dynamic programming for string alignment (either DNA or proteins) with minimal edits, counting and analysis of fixed-length strings (k-mers), and a variety of graph and sparse matrix methods.

ExaBiome's challenge problem is to demonstrate a high-quality assembly or set of assemblies on at least 50 TB of environmental data (reads) that runs across a full-exascale machine. The intent is to use a scientifically interesting environmental sample that may include multiple temporal or spatial samples, which will be processed as a single assembly using complete sequence data. In contrast, current state-of-the-art assembly pipelines are forced to use subsampling when datasets get large, which limits researchers' ability to assemble rare, low-coverage species, and confusing duplications of genomes can result. Furthermore, assembling data across both time and spatial scales will not only enhance the assembly quality but could also reveal functions that otherwise would remain hidden. Addressing this challenge problem will demonstrate a first-in-class science capability using the power of exascale computing combined with novel graph algorithms. This project is expected to provide many potential beneficial science impacts, such as enhancing understanding of microbial functions that can aid in environmental remediation, food production, and medical research. Given the growth of genomic data, a scientifically interesting 50 TB environmental sample should be available by 2022 and is expected to be large enough to fully utilize an exascale machine. However, the challenge problem could also use synthetic data with environmental characteristics or an ensemble assembly of multiple independent environmental datasets. It may also use short reads, long reads, or a hybrid of the two.



PI: Katherine Yelick, Lawrence Berkeley National Laboratory

Collaborators: Lawrence Berkeley National Laboratory, Joint Genome Institute, Los Alamos National Laboratory

Progress to date

- Demonstrated scalable HipMer and MetaHipMer peformance on over 1,000 nodes.
- Implemented overlap/aligner for long reads.
- Completed assembly of a 3 TB large soil data set, the largest metagenome ever assembled. Overall the computation required approximately 4.5 hours on 1,024 Cori KNL nodes, in addition to an hour on a single node (for the memory-intensive scaffolding bottleneck). Memory issues are being addressed by a new scaffolding version.

The ExaBiome project is providing exascale solutions for the assembly and analysis of metagenomic data that will address both current and future data processing needs in bioinformatics.