A centralized and scalable infrastructure approach to support next-generation sequencing at the National Institute of Allergy and Infectious Diseases

Nick Weber, Vivek Gopalan, Mariam Quiñones, Hugo Hernandez, Robert Reed, Kim Kassing, Yentram Huyen, Michael Tartakovksy
National Institute of Allergy and Infectious Diseases, Office of Cyber Infrastructure and Computational Biology, Bethesda, MD 20892, USA.

Recent advances in the “next generation” of sequencing technologies have enabled high-throughput sequencing to expand beyond large specialized facilities and into individual research labs. Improved chemistries, more powerful software, and parallel sequencing capabilities have led to the creation of many terabytes of data per instrument per year that will serve as the basis for diverse genomic research. Because of the massive amounts of data to be managed, many researchers require assistance from IT experts and bioinformaticians in order to store, transfer, process, and analyze all the data generated in their labs. The Office of Cyber Infrastructure and Computational Biology (OCICB) at the National Institute of Allergy and Infectious Diseases (NIAID) has developed a centralized and scalable infrastructure to support next-generation sequencing (NGS) efforts across the Institute. Primary goals of this approach are to standardize practices for data management and storage and to capitalize on the efficiencies and cost savings of a shared high-performance computing infrastructure.

The OCICB’s Operations and Engineering Branch (OEB) has made several significant investments to support NGS research, including improvements in the NIAID network, in data storage and processing hardware, and in the personnel required to build and maintain this infrastructure. Specific upgrades include the following:

- Expansion of network bandwidth from 1 to 10 gigabits per second to support increased network traffic between NIAID research labs and the NIAID Data Center
- Construction of a high-speed and highly-dense enterprise storage system, originally built at 300-terabyte capacity but rapidly scalable to up to 1.2 petabytes
- Creation of a high-performance Linux computing cluster hosting many third-party applications that enables efficient data processing on a scalable and high-memory pool of resources
- Deployment of a localized mirror of the UCSC Genome Browser for rapid data visualization and sharing

In addition to these upgrades, the OCICB’s Bioinformatics and Computational Biosciences Branch (BCBB) will provide bioinformatics collaboration and support to researchers. Specific resources that will be provided include the following:

- End-to-end laboratory information management system (LIMS) to support sample preparation and tracking, task assignment, interaction with the instrument, downstream analysis and custom pipelines between applications, data sharing, and data publication/visualization
- Training on the use of bioinformatics applications and development of custom workflows and application pipelines to streamline data analysis
- Collaboration on the data integration, analysis, and annotation and publication processes

Some policy decisions have yet to be made, including formalizing procedures for long-term data retention as well as balancing data privacy and security requirements while concurrently facilitating data sharing and publication. Nevertheless, NIAID’s centralized approach highlights the need for a cooperative partnership between bench researchers, computational scientists, and IT professionals in order to advance modern scientific exploration and discovery.