INNOVATING CROSS-DOMAIN SOLUTIONS TO DETECT EMERGING BIOLOGICAL THREATS

A Framework For Standardization Of Multi-omics Data Streams Between The Doe Joint Genome Institute And National Microbiome Data Collaborative To Support Fair Data Generation

Kjiersten Fagnan Lawrence Berkeley National Laboratory **Chris Beecroft** Lawrence Berkeley National Laboratory **Chuck Parker** Lawrence Berkeley National Laboratory **Eduardo Lee** Lawrence Berkeley National Laboratory

Background
Many fields of science, including bioscience research, could be revolutionized by advances in data science, machine learning (ML), and artificial intelligence (AI). Leveraging these technologies requires the availability of large, standardized data sets for training and benchmarking. Thus, solving challenges of biological data standardization, normalization, reproducibility, and reliability are key to enable application of these technologies. Data standardization can be loosely described as converting/integrating data into a standard format that computers can easily use. It also enables cross-organizational data integration, reliability, and reuse through rich metadata. Use of standardized data can improve data quality, reduce costs, improve reusability by the broader community, facilitate the development and integration of tools, and make it feasible to develop validation/evaluation datasets for multiple projects.

Purpose
The DOE Joint Genome Institute's Archive and Metadata Organizer (JAMO) is the platform that has been developed to specifically enable data integration and standardization for sequencing information. JAMO is a flexible archiving system that allows suitable metadata (e.g., project identifier(s), data type, software versioning) to be attached to a file as it is being deposited into the JAMO system in a templated manner. JAMO plays a critical role in enabling open access to JGI data by linking it to published ontologies. This framework created an excellent opportunity to build on a well-developed and exercised platform specifically focused on sequencing data that was expanded to manage multi-omics data generated by the National Microbiome Data Collaborative (NMDC). This paper contains a description of the framework infrastructure and reuse by NMDC.

Results and Conclusions
JAMO is the backbone for making JGI data FAIRer. In addition to the templated, standardized metadata that are critical for powerful search, JAMO maintains persistent, globally unique identifiers for each record so a user can reference the same records at any point in the future. Metadata in the JAMO system is woven together to power queries driving results from the JGI's Data Citation Explorer (DCE), a system that identifies reuse of JGI data in publications that did not cite the JGI. This happens when JGI data is downloaded from a national repository like the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA). As these data citations are discovered, the citations are added to JAMO and the knowledge graph of multi-omics metadata connections in DCE becomes more powerful. JAMO manages more than 16 million files, 3 billion pieces of metadata, and 16 petabytes of standardized multi-omics data. JGI is in the process of making this software infrastructure open source and available to the broader community.

Relevance and Impact
JAMO is a framework that provides a means for curation and aggregation of large data sets required for use in ML/AI-based approaches to synthesize information across studies, identify connections between disparate resources linked by global identifiers (NMDC, JGI, NCBI), and quickly identify biosecurity threats. This foundational software infrastructure can be reused or leveraged as a model by other research organizations and teams to support FAIR data management and information exchange.

The work was conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231.