Multi-omics workflows to support data integration for the National Microbiome Data Collaborative

Bin Hu[4]* (bhu@lanl.gov), Faiza Ahmed[2], Anubhav[3], Jeffrey Baumes[2], Jonathan Beezley[2], Mark Borkum[3], Lisa Bramer[3], Shane Canon[1], Patrick Chain[4], Danielle Christianson[1], Yuri Corilo[3], Karen Davenport[4], Brandon Davis[2], Meghan Drake[5], William Duncan[1], Kjiersten Fagnan[1], Mark Flynn[4], Marcel Huntemann[1], Julia Kelliher[4], Sonya Lebedeva[1], Po-E Li[4], Mary Lipton[3], Chien-Chi Lo[4], Douglas Mans[3], Stanton Martin[5], Lee Ann McCue[3], David Millard[3], Kayd Miller[1], Nigel Mouncey[1], Chris Mungall[1], Paul Piehowski[3], Elais Player Jackson[4], Anastasiya Prymolenna[3], Samuel Purvine[3], TBK Reddy[1], Rachel Richardson[3], Migun Shakya[4], Montana Smith[3], Jagadish Chandrabose Sundaramurthi[1], Deepak Unni[1], Pajau Vangay[1], Bruce Wilson[5], Donny Winston[6], Elisha Wood-Charlson[1], Yan Xu[4], **Emiley Eloe-Fadrosh**[1]

[1] Lawrence Berkeley National Laboratory, Berkeley, CA; [2] Kitware, Clifton Park, NY; [3] Pacific Northwest National Laboratory, Richland, WA; [4] Los Alamos National Laboratory, Los Alamos, NM; [5] Oak Ridge National Laboratory, Oak Ridge, TN; [6] Polyneme LLC, New York, NY

Project Goals: The National Microbiome Data Collaborative (NMDC) is a pilot initiative launched to support microbiome data exploration and discovery through a collaborative, integrative science gateway. With a community-centered design approach, the NMDC team is building an open-source, integrated data science ecosystem that leverages existing data standards, and data resources and infrastructure within the DOE complex.

Abstract

Standardized omics workflows drive the analysis of raw omics data and ensures the data stored in the National Microbiome Data Collaborative data portal[1] are processed in a uniform fashion and comparable across studies. The NMDC source code repository[2] offers workflows to perform Illumina paired-end reads quality control, metagenomic and metatranscriptomic, metabolomic and metaproteomic analysis. These best practice workflows are developed on top of decades of omics analysis experience gathered from participating institutions, with all computing environment dependencies removed, and coded in the workflow description language (WDL[2]). They are packaged as software containers[3] and documented[4] to enable microbiome researchers to install and run workflows locally, to understand the tools and uses for each workflow, and to further allow local workflow improvements or customisations to meet their specific requirements. By leveraging these workflows, researchers can analyze their data by themselves and expect the same results as if their data were processed by the NMDC portal. A web platform (NMDC EDGE) running these workflows interactively will be provided through the next version of the EDGE bioinformatics suite and similar integration is planned for the DOE KnoweldgeBase (KBase) in the future.

References

[1] https://data.microbiomedata.org/

[2] https://github.com/microbiomedata/

[3] https://www.commonwl.org

[4] https://hub.docker.com/u/microbiomedata

[5] https://nmdc-workflow-documentation.readthedocs.io/en/latest/

Funding statement

This work is supported by the Genomic Science Program in the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research (BER) under contract numbers DE-AC02-05CH11231 (LBNL), 89233218CNA000001 (LANL), DE-AC05-00OR22725 (ORNL), and DE-AC05-76RL01830 (PNNL).