On the way to research objects for environmental genomics (or metagenomics)

Folker Meyer
MCS
Argonne National Laboratory
Argonne, IL, USA
folker@anl.gov

Robert D. Finn
Sequence Families Team,
EMBL-EBI European Bioinformatics Institute,
Cambridge, UK
rdf@ebi.ac.uk

Wolfgang Gerlach
Data Science and Learning
Argonne National Laboratory
Argonne, IL, USA
wgerlach@mcs.anl.gov

Alex Mitchell
Sequence Families Team,
EMBL-EBI European Bioinformatics Institute,
Cambridge, UK
mitchell@ebi.ac.uk

Travis Harrison
Data Science and Learning
Argonne National Laboratory
Argonne, IL, USA
tharriso@mcs.anl.gov

Andreas Wilke
Data Science and Learning
Argonne National Laboratory
Argonne, IL, USA
wilke@anl.gov

The MG-RAST portal [Meyer] and its European sister project MGnify [Mitchell]] at the European Bioinformatics Institute (EMBL-EBI) provide metagenome analysis services to a large, international community of scientists. The systems capture metadata about each data set according to the standards of the Genomics Standards Consortium (GSC) [Field] and both have more recently begun to convert their workflows to Common Workflow Language [CWL] format.

Both metagenome data and computation with metagenomes are expensive [Thomas], significant degrees of freedom exists for the computational analysis underscoring the need for reproducibility in the field of environmental sequence analysis. While existing initiatives are attempting to benchmark different computational approaches [Sczyrba], it is vital for researchers to understand the provenance of information derived from metagenomes. Our CWL formatted workflows allow rapid comparison of the two pipelines, with CWL described tools being reused to from other, related workflows for the analysis of marine eukaryotic transcriptomics.

MG-RAST has captured metadata about the data objects using GSC standards for several years and is exporting those via RESTful APIs [Wilke] and MGnify [Mitchell]. Together we expect to use Research Objects to export provenance information as part of our APIs. We are now working towards domain specific profiles, evaluating CWLProv [Khan] to identify any community specific extensions that might be needed for the Microbiome research community.

Works Cited

"Common Workflow Language, v1.0." *PLOS Biology*, Public Library of Science, doi.org/10.6084/m9.figshare.3115156.v2.

Field, Dawn, et al. "The Genomic Standards Consortium." *PLoS Biology*, vol. 9, no. 6, 2011, doi:10.1371/journal.pbio.1001088.

Khan, Farah Zaib, et al. "CWLProv - Interoperable Retrospective Provenance Capture and Its Challenges." *PLOS Biology*, Public Library of Science, 27 Mar. 2018, doi.org/10.5281/zenodo.1208477.

Meyer, F, et al. "The Metagenomics RAST Server – a Public Resource for the Automatic Phylogenetic and Functional Analysis of Metagenomes." *BMC Bioinformatics*, vol. 9, no. 1, 2008, p. 386., doi:10.1186/1471-2105-9-386.

Mitchell, Alex L, et al. "EBI Metagenomics in 2017: Enriching the Analysis of Microbial Communities, from Sequence Reads to Assemblies." *Nucleic Acids Research*, vol. 46, no. D1, 2017, doi:10.1093/nar/gkx967.

Sczyrba, Alexander, et al. "Critical Assessment of Metagenome Interpretation—a Benchmark of Metagenomics Software." *Nature Methods*, vol. 14, no. 11, Feb. 2017, pp. 1063–1071., doi:10.1038/nmeth.4458.

Thomas, Torsten, et al. "Metagenomics - a Guide from Sampling to Data Analysis." *Microbial Informatics and Experimentation*, vol. 2, no. 1, 2012, p. 3., doi:10.1186/2042-5783-2-3.

Wilke, Andreas, et al. "A RESTful API for Accessing Microbial Community Data for MG-RAST." *PLoS Computational Biology*, vol. 11, no. 1, Aug. 2015, doi:10.1371/journal.pcbi.1004008.