Sample Size Does Matter: Scaling Up Analysis in Galaxy with Metagenomics

Daniel Blankenberg[1, 2], Sarah Carnahan-Craig[3], and the Galaxy Team[2]
[1] Department of Biochemistry and Molecular Biology, Penn State University, University Park, PA 16802, USA. Email: dan@bx.psu.edu
[2] https://galaxyproject.org.
[3] Department of Biology, Penn State University, University Park, PA 16802, USA.

Project Website: galaxyproject.org **License**: Academic Free License version 3.0
Source Code: github.com/galaxyproject/galaxy

Galaxy (http://galaxyproject.org) is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research. Galaxy makes bioinformatics analyses accessible to users lacking programming experience by enabling them to easily specify parameters for running tools and workflows. Analyses are made transparent by allowing users simple access to share and publish analyses via the web and create Pages, interactive, web-based documents that describe a complete analysis.

Metagenomics provides an exciting opportunity to begin to explore large-scale multiple sample analysis with Galaxy. As part of an obesity study, we have obtained over 400 buccal and stool samples from mother-child pairs. These samples have been subjected to 16S RNA extraction and sequencing on a MiSeq instrument. While sequencing 400 samples is no small feat, once generated, the data analysis reveals itself as crippling bottleneck.

Galaxy provides researchers with a vast quantity of tools and methods to analyze a wide-array of data, and makes connecting any number of tools together easy via Workflows. Although running a workflow individually over a handful of samples is approachable, how does one deal with 10, 20, or even 100 samples without becoming frustrated, introducing errors, breaking their mouse, or falling back to writing an API script? While Dataset Collection functionality provides a significant portion of a solution to this problem, there are still major hurdles that need to be overcome before Galaxy is usable for large multiple sample analysis.

Here we describe a generalizable metagenomic pipeline as implemented within Galaxy that is able to handle the simultaneous analysis of over 5,000 Human Microbiome Project samples. In addition to integrating a number of third-party algorithms and toolsets, some requiring the creation of upstream fixes and enhancements, we have developed new tools and approaches for dealing with large collections of data. Furthermore, we discuss the problems encountered using Galaxy at a large-scale, what has been done to overcome these issues, as well as initial results.