Demystifying the Interoperability of Disparate Genomic Resources

Daniel Blankenberg[1, 2] 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] http://galaxyproject.org.

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.

Galaxy provides experimental biologist access to powerful analysis infrastructure through the web. Within Galaxy, users are able to upload their own data, graphically execute command-line analysis tools, and interactively build visualizations on their results. In many circumstances, a user is able to conduct their entire analysis without leaving the Galaxy application. However, this is not always the case, as users may find themselves in a situation where they want to utilize data from an external data warehouse or where they may want to continue an analysis at an external resource.

Although Galaxy strives to provide an all-inclusive analysis platform, there are times when external resources provide functionality that is not available within, or is superior to, those embedded directly as a part of Galaxy. These apparent weaknesses actually highlight some of Galaxy's strongest features. Galaxy embraces external resources by providing frameworks for retrieving and sending data seamlessly through its graphical interface. This is more than simply transferring files between two web-servers.

When retrieving data from external resources, file content is only one part of the puzzle. There is often important metadata associated with the datasets, such as genome build identifiers, formats, sample information, column assignments, versions, etc., that are required for the datasets to be useful. Galaxy enables this through the use of Data Source tools, where the framework has been recently enhanced to allow external resources to send multiple files at a time and to provide extensive amounts of metadata. Galaxy currently supports several external data resources, including UCSC table browser, Biomart, InterMine, European Nucleotide Archive, and GenomeSpace. And adding more is a straightforward process.

When sending datasets to an external resource, there are typically two different methods available. In the first method, a standard Galaxy tool can be defined that allows a user to pick datasets from their history, configure any additional options, and then click *Execute* to send the data to the external resource. Data export tools will typically create a new HTML Galaxy History item that contains the results and hyperlinks to allow the user to transition to the external resource. The second method is generally used for enabling external visualization tools. Using the external display application framework, link-outs to external resources, such as UCSC Genome Browser, Integrative Genome Viewer (IGV), and GBrowse, are embedded directly within the dataset preview in the user's history. The user can simply click the link under their dataset and will be forwarded directly to the external application along with their data. In cases when the external display requires different formats or additional index files, such as viewing a VCF file within IGV (i.e. bgziped and indexed with tabix), standard Galaxy converter tools can be automatically utilized by the framework.

Here, we demonstrate a typical NGS analysis where we provide our own data, retrieve data from an external data source, perform an analysis on these data within Galaxy and then send our data to additional external resources for further analysis and visualization. We then explore the basic steps that were needed to enable these external interactions within Galaxy.