Increasing the utility of Galaxy workflows.

John Chilton[1] and The Galaxy Team
[1]Department of Biochemistry and Molecular Biology, Penn State University, University Park, PA 16801, US
jmchilton@bx.psu.edu

Project: http://galaxyproject.org/   Code: http://github.com/galaxyproject/galaxy

Galaxy is a popular data analysis platform that is most often used to integrate diverse command-line utilities into a consistent and intuitive web-based interface. A long standing selling point of Galaxy has been that it allows researchers to extract sample analysis histories out into reusable workflows and build such workflows de novo. Despite the popularity of this feature, the kinds of workflows that could be expressed by Galaxy have had critical limitations. Two of the most glaring of these are that Galaxy workflows have required a fixed number of inputs and the workflow engine planned out every job right at submission time (workflows would cause a series of jobs to queue up in Galaxy - but there was no real workflow engine that could alter the structure of this computation over time). Many relatively basic analyses in bioinformatics require running a variable number of inputs across identical processing steps (“mapping”) and then combining or collecting these results into a merged output (“reducing). Likewise - pausing workflows, splitting inputs up into multiple datasets, and conditionals all require the re-evaluation of workflows overtime. This presentation will discuss how we have started addressing these limitations. In particular we will present dataset collections and a real, pluggable Galaxy workflow subsystem - together these features address the limitations described above and vastly increase the expressiveness of Galaxy workflows.

Galaxy dataset collections are powerful way to group collections into potentially nested hierarchies of lists and pairs of datasets. Existing Galaxy tools can be used without modification to “map” operations across dataset collections to produce new collections with sample information maintained. Likewise tools that consume many datasets can be readily used to “reduce” these collections. For newly developed tools - a wide range of extensions to Galaxy tooling format exist to consume and produce dataset collections. In addition to presenting these additions to Galaxy, extensions to the workflow system to tie together these analyses and innovative UI elements such as the paired list dataset collection builder will be presented.

Specific biologically relevant examples to highlight the power of dataset collections and the new workflow engine will be presented. These will include an RNA-seq workflow based on the tuxedo suite of tools that can process any number of samples and a workflow that exploits the ability to output collections to achieve greater parallelization than was previously possible.

These extensions are powerful new features that greatly enhance the expressivity of Galaxy workflows, but much work remains to do be done. A road map for the future of Galaxy workflows will be laid out - including conditionals, iteration, and more flexible connections between steps (e.g. mapping output metadata to input parameters for instance), etc....