Large scale analyses with improved dataset collections

Marius van den Beek\textsuperscript{1}, John Chilton\textsuperscript{2}, Nicola Soranzo\textsuperscript{3}, Anil S. Thanki\textsuperscript{3}, Martin Čech\textsuperscript{2}, and the Galaxy Team

\textsuperscript{1}Institut Curie, 26 rue d’Ulm, F-75248 Paris, France.
\textsuperscript{2}Department of Biochemistry and Molecular Biology, Penn State University, University Park PA USA
\textsuperscript{3}Earlham Institute (EI), Norwich Research Park, Norwich NR4 7UZ, UK.
m.vandenbeek@gmail.com & jmchilton@gmail.com

In 2014, the Galaxy team rolled out dataset collections for scaling analyses to multiple datasets. Since then the Galaxy team and community worked to increase the scale and complexity of the analyses possible with dataset collections demonstrating that it is possible to use collection for analysis of thousands of samples generating terabytes of data and hundreds of thousands of intermediate datasets from the Galaxy user interface without any requirement for knowledge of scripting or command lines.

Over the last year, we have revisited nearly every aspect of the user interface as it relates to collections trying to make such large-scale analyses not just possible in Galaxy, but easy and intuitive, truly accessible regardless of computational expertise.

Collections now work with Galaxy’s uploader, dataset libraries, workflow actions, display applications, tags, and more tools and workflows than ever before. Galaxy has added powerful new collection operations that allow managing the structure and flow of collections through an analysis without needing to write a script or line of code. Collection state is now rendered in the history panel - providing condensed, real-time information about potentially hundreds of thousands of jobs.

Previously the Galaxy user interface was particularly lacking in two areas: dealing with nested collections, and loading large numbers of files into Galaxy. We will highlight recent work that addresses both of these problems. We will demonstrate a user interface for describing rules for the creation and manipulation of collections. The interface builds a completely reproducible program for the manipulation of metadata using a simple tabular-centric domain-specific language, but our belief is that this interface feels as simple and intuitive as manipulating data in a spreadsheet and requires no particular computational expertise. Using this widget, we will demonstrate the creation of collections with rich metadata from sample sheets or various online data repositories. We also demonstrate how this same technique can be used to reorganize existing collections from metadata during the middle of an analysis to leverage Galaxy’s powerful options for mapping and reducing effectively and flexibly. Finally we will outline strategies for an optimal analysis of a large number of small datasets or a small number of very large datasets.

Throughout the presentation, we will demonstrate these enhancements using real world examples of best practice analyses in Galaxy.