Reproducible and automated processing in high-throughput NGS facilities

As the rate of samples to process in high-throughput sequencing facilities increases, performing and tracking the center's operations becomes increasingly difficult, costly and error prone, while processing the massive amounts of data poses significant computational challenges.

We present our ongoing work to accelerate, automate and track all data-related procedures at the CRS4 Sequencing Platform by integrating Galaxy with other state-of-the-art processing technologies, such as Hadoop, OMERO and iRODS.

In our model, data processing pipelines are implemented as one or more Galaxy workflows. Through our integration work, in addition to conventional tools Galaxy is able to drive high-performance Hadoop-based processing tools. With all workflows, Galaxy tracks the processing steps applied to data through its histories; as data sets are generated, these histories are extracted and stored into our OMERO.biobank, thus documenting the data and ensuring reproducibility. The data itself, on the other hand, is committed to iRODS, hence providing a single file repository, independent from the storage infrastructure. A custom "automator" daemon is the final component required to drive the system. It launches and monitors Galaxy workflows, links workflows to each other – e.g., execute sample-based workflows after a flowcell-based workflow – and passes information between components – i.e., saves data sets and histories in OMERO.biobank, commits files to iRODS, etc.

Currently, the system is in its testing phase and is on schedule to be in production at CRS4 by May 2013. In addition, future extensions will allow it to be used to process data from other sources, such as mass spectrometers and digital microscopes.