We present the key software components of the automated infrastructure (see Fig.1) for tracking all data-related procedures and clinical information, deployed at the CRS4 high-throughput sequencing platform, which is the largest in Italy (10TBase of sequencing data per month).

The CRS4 automation layer takes care of processing of the data produced by the sequencing machines, orchestrating format conversion, demultiplexing, quality checks and preparing the data for on-site storage or delivery in order to minimize human intervention and increase the reliability.

The full description of biomedical datasets, the graph of dependencies between data sets, and the computational procedures that derived them is provided by the core components of the infrastructure OMERO.biobank (see Fig.3-4), which is a robust, extensible and scalable traceability biodata management system that provides the mechanism to store, query and retrieve metadata.

In order to simplify and manage access to the data and to provide a single logical namespace for all data sets, CRS4 adopted iRODS as a front end to his large scale heterogeneous storage system. iRODS implements data virtualization, allowing access to distributed storage assets under a unified namespace.

Finally, our workflow management system is based on Galaxy, an open platform for reproducible data-intensive computational analysis that provides a standard way to encapsulate computational tools and data sets in a graphical user interface (GUI), together with a mechanism to keep track of execution history.

The integration of open-source tools, either internally developed or publicly available, allowed us to establish a framework that provides automatic sequencing data ingestion, data provenance, custom pipeline tuning and versioning, flexible and reliable data virtualization. This infrastructure is currently involved in a number of clinical studies in cooperation with leading research institutions in Italy and abroad.

Reference:

Contacts: gianmauro.cuccuru@crs4.it
Take home: https://goo.gl/Ty82qi