A robust and extensible framework for tracking biodata provenance, analysis workflows and results

Gianmauro Cuccuru, Rossano Atzeni, Luca Lianas, Ricardo Medda, Luca Pireddu, Paolo Uva, Ilenia Zara, Gianluigi Zanetti, Giorgio Fotia
Centro di Ricerca, Sviluppo e Studi Superiori in Sardegna (CRS4), Pula, Cagliari

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).

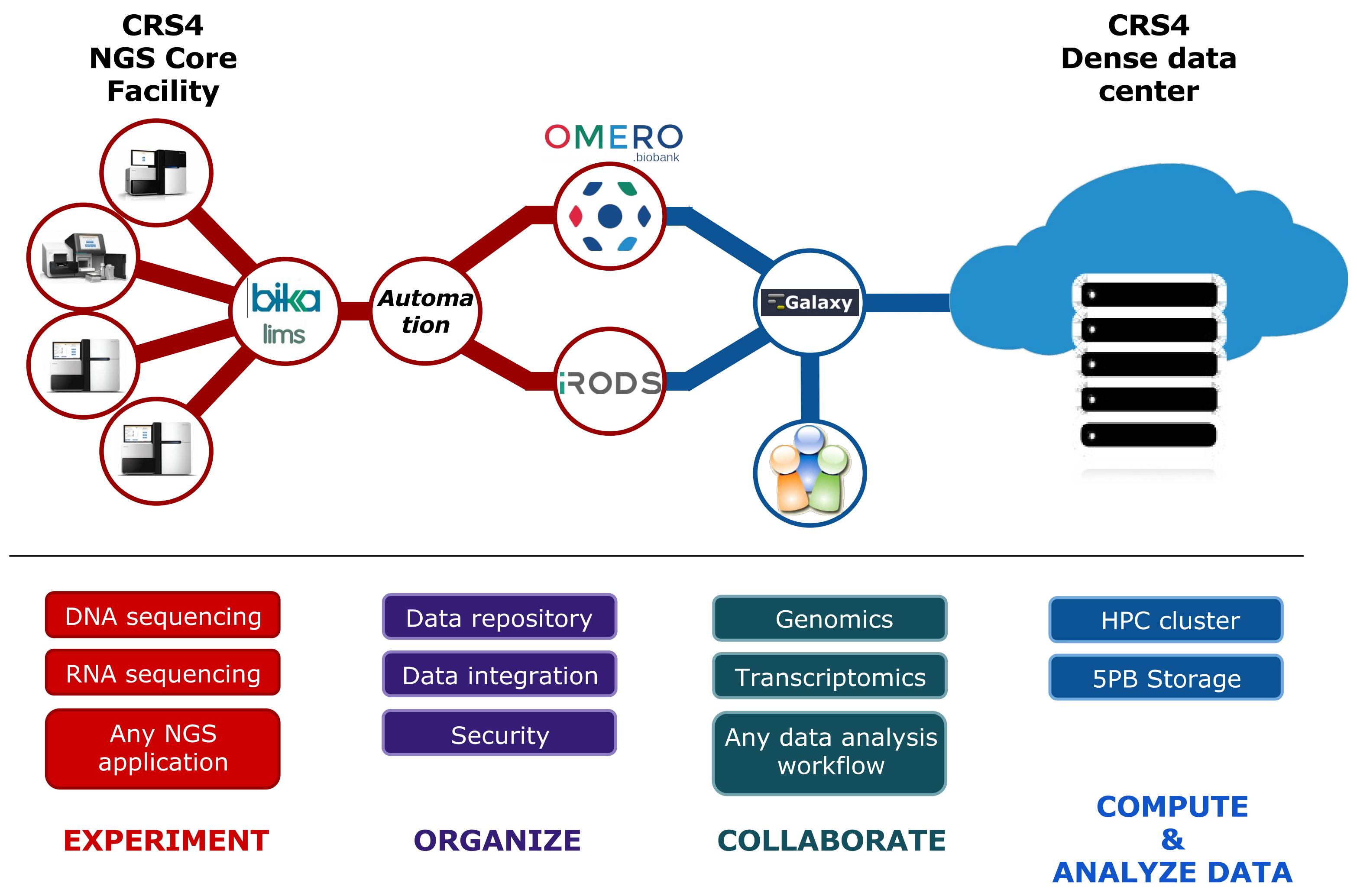
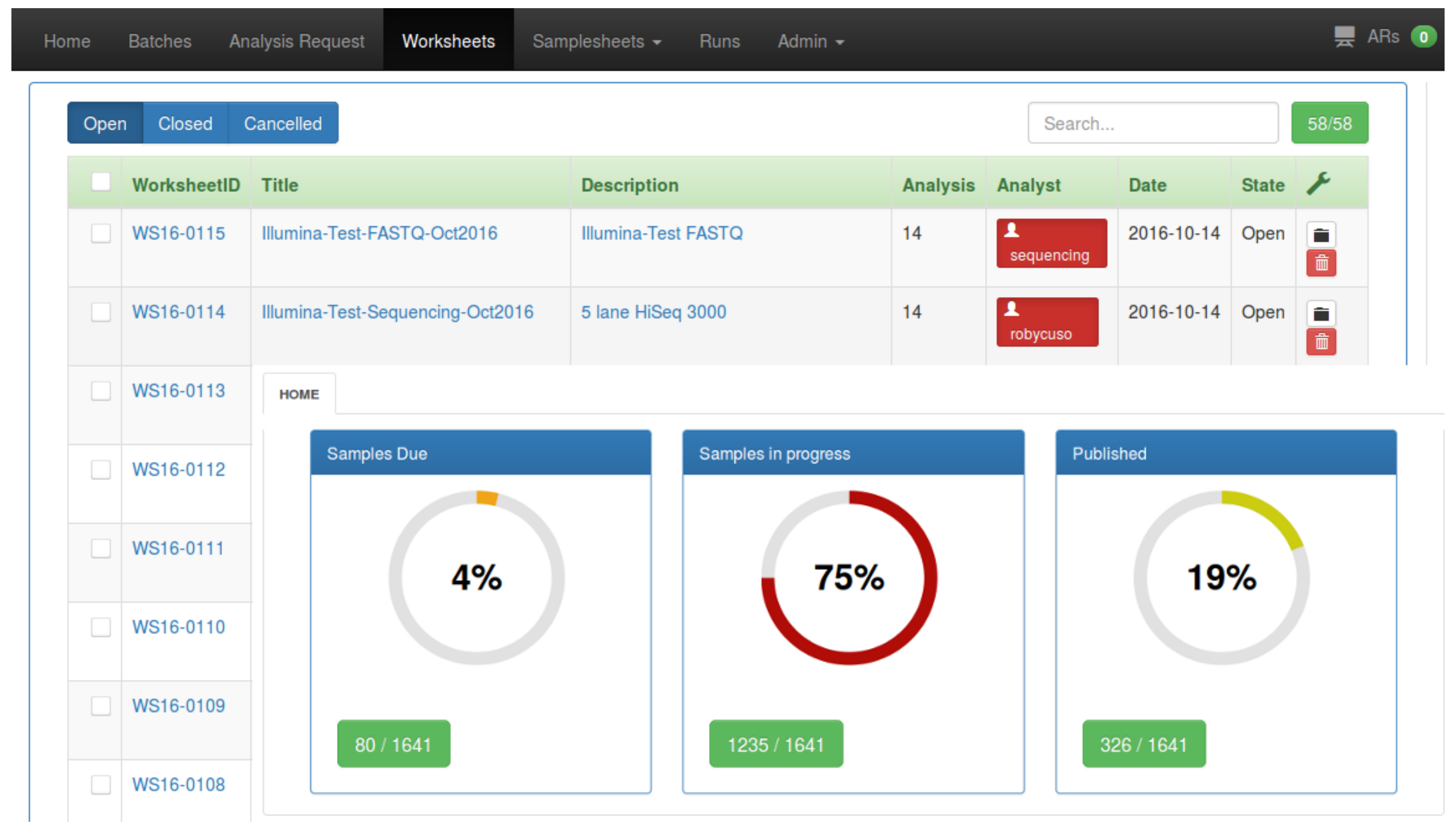


Figure 1. Main components of the automated CRS4 infrastructure to support end-to-end analysis of sequencing data (Cuccuru et al., 2014).

Wet-lab activities are tracked using the open source, community based software BIKA-LIMS, exposed to the wet-lab technicians through a REST web-service and a custom AngularJS interface (see Fig.2).

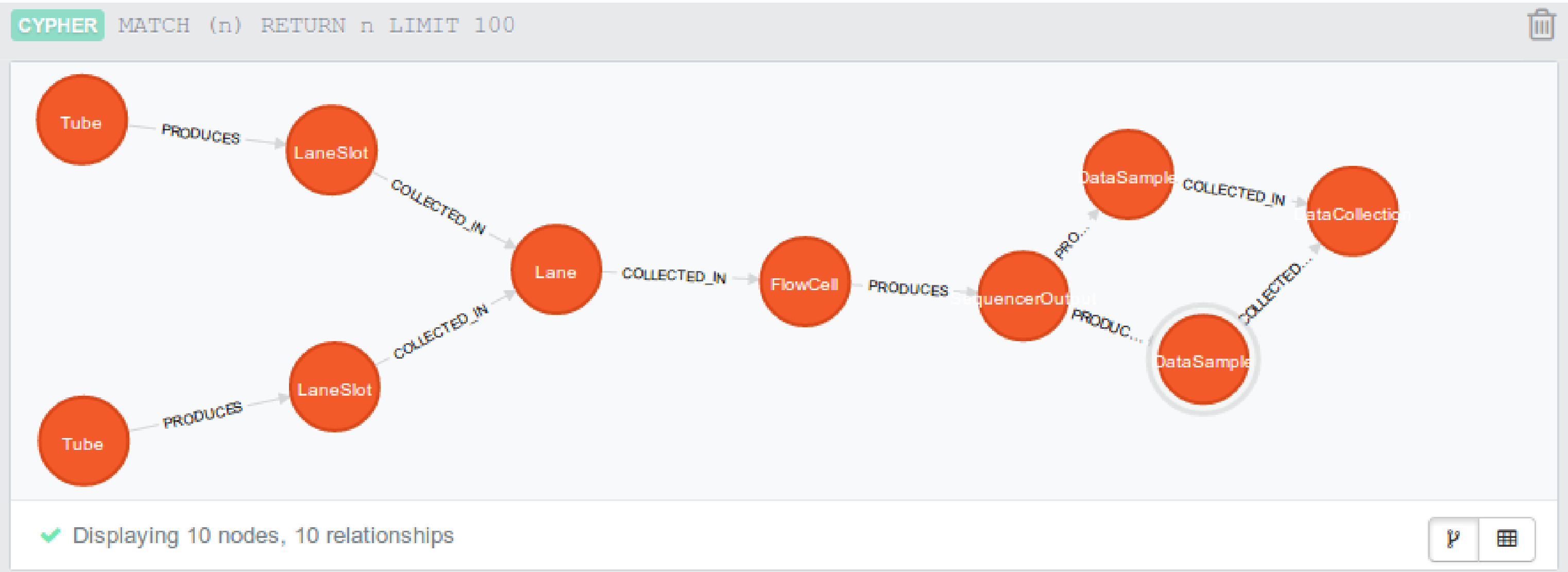


	WorksheetID	Title	Description	Analysis	Analyst	Date	State
□	WS16-0115	Illumina-Test-FASTQ-Oct2016	Illumina-Test FASTQ	14	sequencing	2016-10-14	Open
□	WS16-0114	Illumina-Test-Sequencing-Oct2016	5 lane HiSeq 3000	14	robycuso	2016-10-14	Open
□	WS16-0113						
□	WS16-0112						
□	WS16-0111						
□	WS16-0110						
□	WS16-0109						
□	WS16-0108						

Figure 2. Screen-shot of BIKA-LIMS custom dashboard.

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.



```
CYPHER MATCH (n) RETURN n LIMIT 100
```

✓ Displaying 10 nodes, 10 relationships

Figure 3. Diagram illustrating relationships between the sequencing objects within OMERO.biobank, from the tube of DNA to the dataset ready for the downstream analyses.

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.

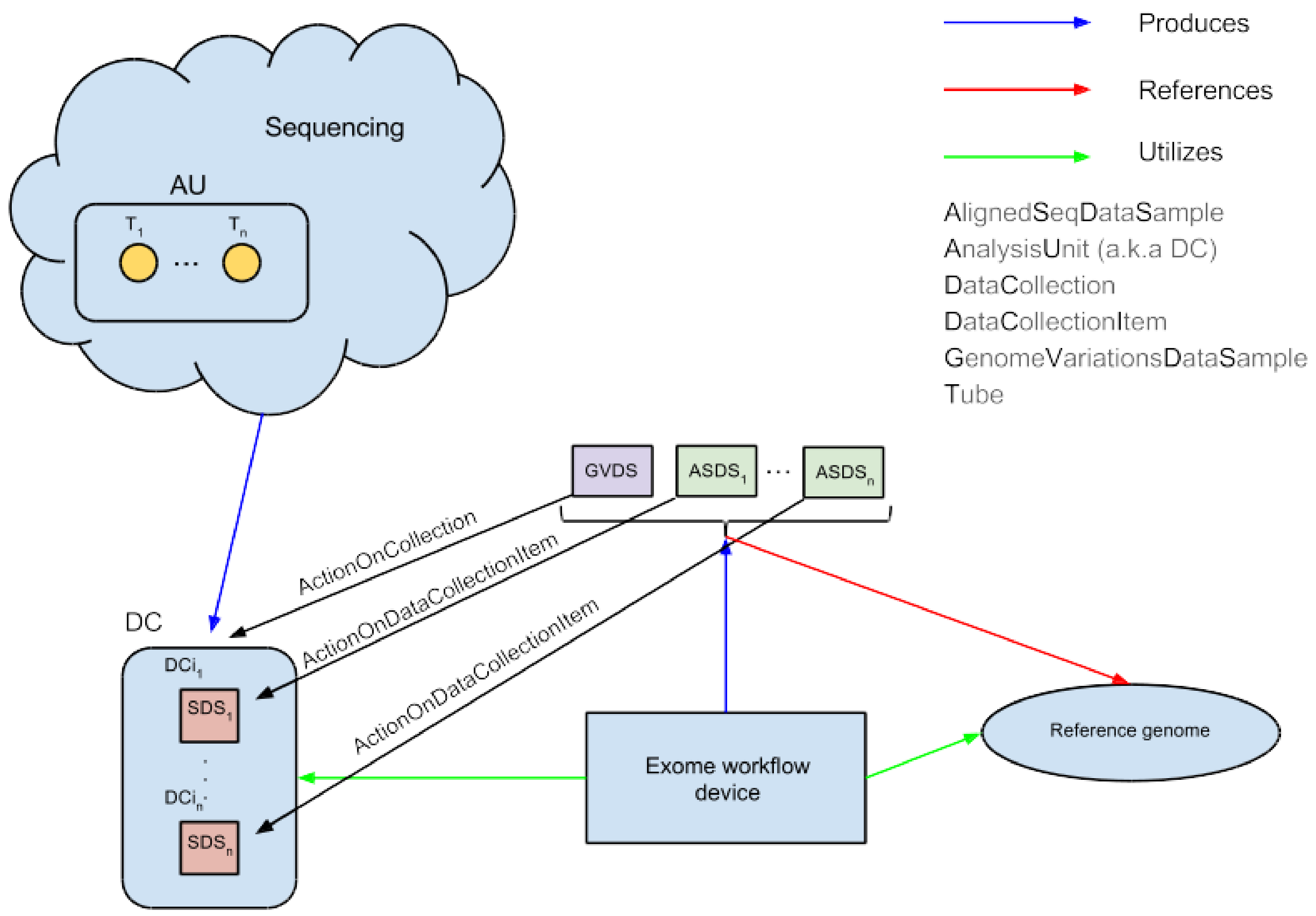


Figure 4. Traceability graph for an exome processing workflow stored within OMERO.biobank.

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:

Cuccuru G. et al. (2014) An Automated Infrastructure to Support High-troughput Bioinformatics. In: Smari, Waleed W. and Zeljkovic V (ed) Proc. IEEE Int. Conf. High Perform. Comput. Simul. (HPCS 2014). IEEE, pp 600–607

Contacts:
gianmauro.cuccuru@crs4.it

Take home:
https://goo.gl/Ty82qi