DBS

DIPARTIMENTO
DI BIOSCIENZE

Laniakea: Shaping and Developing a Cloud-based bioinformatic platform

UNIVERSITÀ DEGLI
STUDI DI MILANO

Pietro Mandreoli[1,3], Marco Antonio Tangaro[1,2], , Matteo Chiara[1,3], Giacinto Donvito[2], Marica Antonacci[2], Graziano Pesole[1,4], Federico Zambelli[1,3]

1. Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies - National Research Council (IBIOM-CNR), Bari, Italy
2. National Institute for Nuclear Physics, Bari Section, Italy
3. Dept. of Biosciences, University of Milan, Italy
4. Dept. of Biosciences, Biotechnologies and Pharmacological Sciences, University of Bari "Aldo Moro", Italy

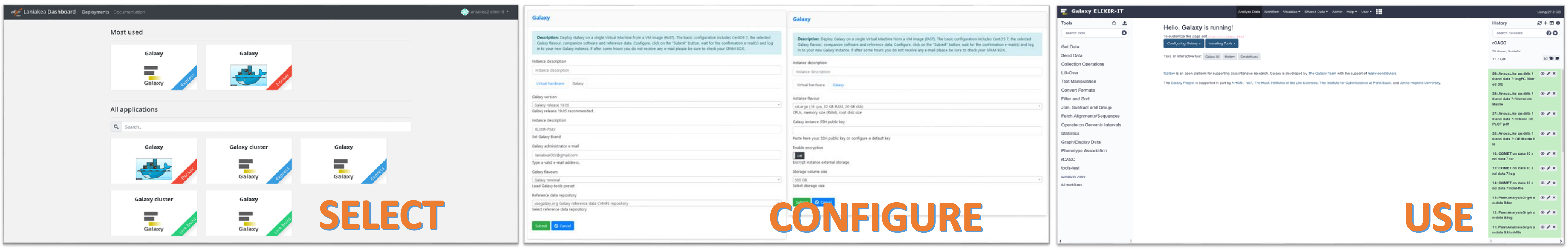
PhD Workshop 2021 UNIMI, 7-8 October

Galaxy is currently the prevailing workflow manager for bioinformatics thanks to its many useful features and a user-friendly interface. While several Galaxy public services are available to researchers, either for general purpose or dedicated to specific research domains, there are still many scenarios where a private Galaxy instance is necessary or preferable, including heavy data analysis workloads, data privacy concerns or specific customization needs.

Cloud computing gives the possibility to the scientific community to have access to state-of-the-art computational resources without having to worry about hardware installation, configuration and maintenance. This characteristic is extremely useful in scientific fields, including life-science , that deal with large amounts of data and need a lot of computational and storage resources to analyse them. Cloud computing technology is a promising solution to this problem allowing small laboratories and even large institutions to have access to adequate IT resources without the need of establishing and maintaining complex and expensive computational infrastructures.

Laniakea is a software framework that facilitates the provisioning of on-demand Galaxy instances as a cloud service over e-infrastructures, by leveraging on the open-source software platform developed by the INDIGO-DataCloud H2020 project, which aimed to make cloud e-infrastructures more accessible by scientific communities.

End-users interact with Laniakea through a web front-end allowing a general setup of a Galaxy instance. The deployment of the virtual hardware and of the Galaxy software ecosystem is subsequently performed by the INDIGO Platform as a Service layer. At the end of the process, the user gains access to a private, production-grade, fully customizable, Galaxy virtual instance.



Improving Laniakea sustainability and maintenance

Galaxy installation and configuration

The Ansible engine is widely used on Cloud environments to orchestrate the installation and configuration of systems and applications. In particular, Laniakea exploits Ansible roles, scripts written using the Ansible automation language to install and configure Galaxy and its tools on the Virtual Machine.
The system has been reworked integrating the official Galaxy Project Ansible roles, used by the community to deploy and configure Galaxy instances. This will allow for faster updates (the GalaxyProject roles are updated at the same time as the release of a new version of Galaxy), making it easier to maintain the Laniakea framework. This development is carried out within the framework of the H2020 EOSC-Pillar project (Work Package 6, Task 6.6.1), which aims at a greater rationalization of the Ansible roles for the Galaxy deployment.

Galaxy deployment strategy

The Laniakea Express system allows a faster Galaxy deployment, since Galaxy, its ancillary software and the bioinformatic tools are packaged in a single Cloud image, easily deployable on any Cloud e-infrastructure. Moreover, this deployment strategy allows us to install Galaxy regardless the online availability of the necessary software, as they are already present in the image and is less time-consuming deployment for Flavors that contain a large number of tools.
As a contraindication, the system is heavy to maintain, as it requires the images to be recreated at each release of Galaxy for any Galaxy tools preset (Galaxy flavor) available in Laniakea.
Therefore, we have completely reshaped the Galaxy Express framework, by improving the image creation, the virtual machine deployment, the bioinformatic tools and workflow installation.
In particular only one image, containing the Galaxy platform, is needed. Then the bioinformatic tools and workflows are restored at the deployment time copying the software packages and system configuration files using a CVMFS storage repository.
The image creation step has been simplified so it can be also automated. The same thing has been done for the tools and workflow installation steps making the whole installation process more flexible and maintainable.

Laniakea sustainability

Laniakea needs to be continuously updated to allow the deployment of new Galaxy releases. Moreover, the more complex and feature-rich the service becomes, the more individual components can break during frequent updates, leading to malfunctions of the system. For this reason we decided to support Laniakea using a Continuous Integration system, allowing Laniakea developers and contributors to merge code on our central GitHub repository, while it is continuously tested.
This system has two main goals: to allow frequent and stable updates executing automatic tests every time a component is changed and to periodically test all Laniakea deployments, tools and workflows.
The system is under testing and exploits different open-source software to check a specific Laniakea component e.g Molecule for Ansible Roles, GitHub actions for docker container, Jenkins for image builds, tools and workflows testing.

Improving Reference Data sharing

In every Galaxy instance the bioinformatics reference data (genomes, genome indexes, variant databases), needed to run specific tools are made available to the users through CernVM Filesystem. The CernVM Filesystem is a central repository that, like a cloud USB stick, hosts a large amount of data and can be mounted on any unix like Virtual Machine, allowing the user of the VM the reference data access, in read-only mode.
Collaborating with the Consortium GARR, we are planning to create a copy (mirror) of the Galaxy community CernVM Filesystem on GARR Cloud. This mirror will allow a more stable and fast reference data sharing on every new Galaxy instance served by Consortium GARR network, through Laniakea or manually.

UPDATED Laniakea

Improving User Experience

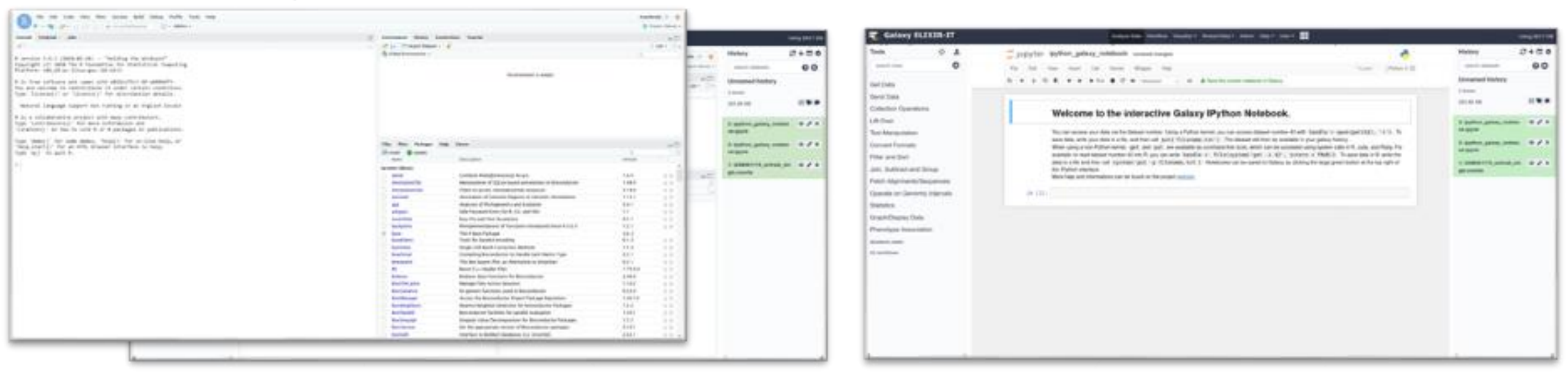
New applications

The Laniakea system was also modified to run two new Applications Rstudio (Interactive Development Enviroment for R) and Rstudio + Seurat (most popular tool for single-cell data analysis). These two applications were requested by Laniakea@ReCaS users for training and development purposes.

The two applications were brought in both in the live build and Express system, are currently under testing phase.

New Galaxy features

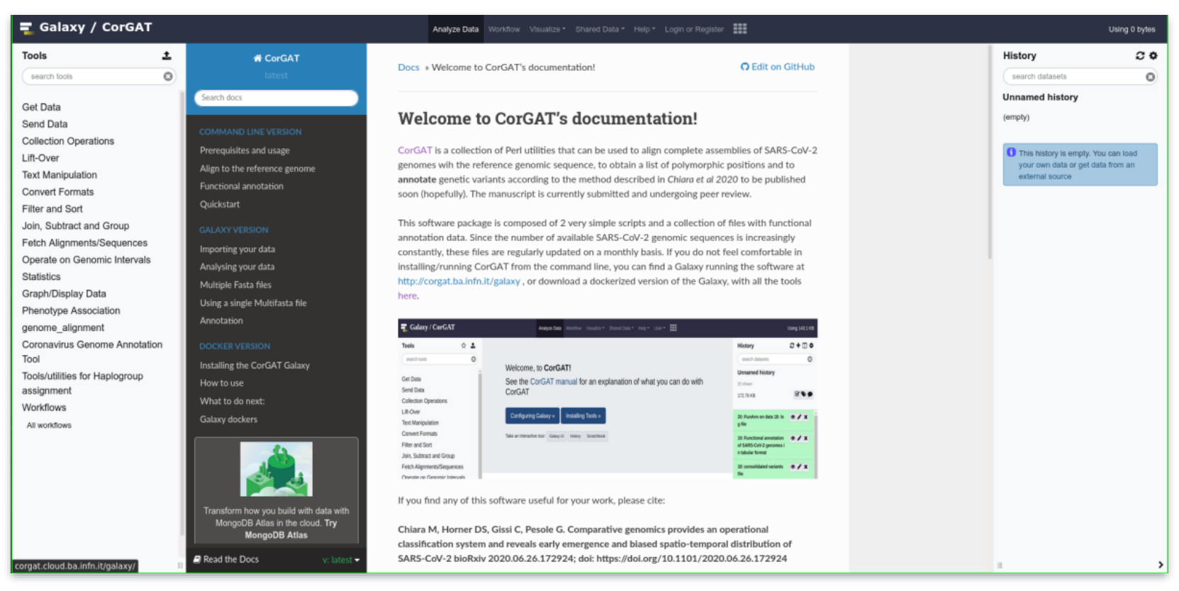
We have also configurated the new Galaxy instances to run Rstudio and Jupiter notebook as integrated application, allowing users to further analyze their data through python and R scripts.



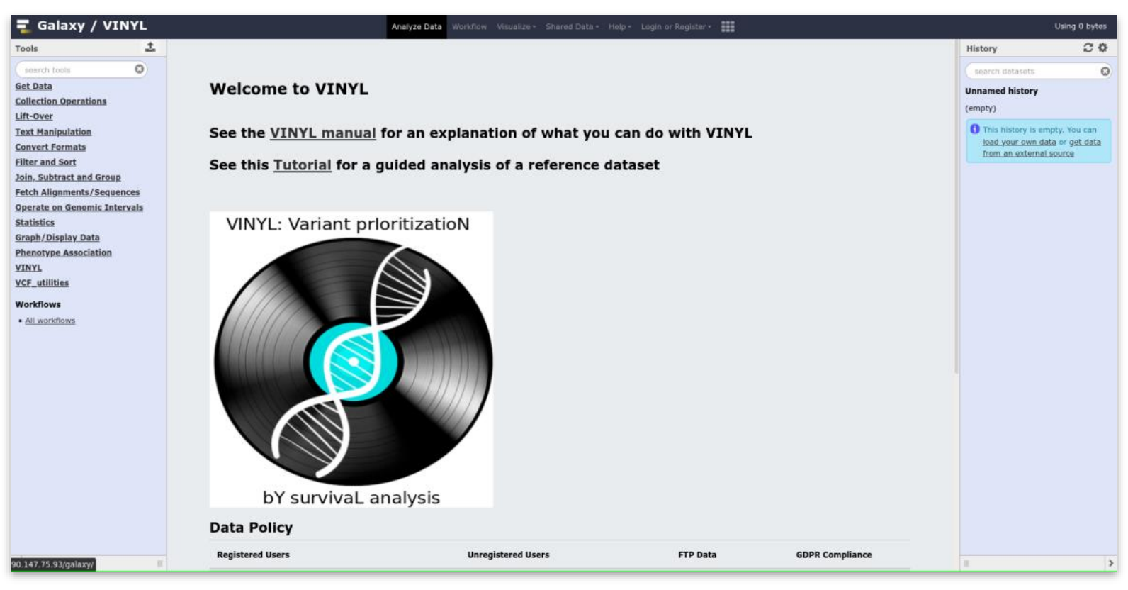
"Laniakea@ReCaS", the first instance of a Laniakea-based service, is managed by ELIXIR-Italy and was officially launched in February 2020. European researchers can request access to Laniakea@ReCaS through an open-ended call for use-cases.

Currently the service is used by 15 research groups and organizations. For a total of 18 Galaxy instances, 113 users, for a total resource budget of 170 vCPUs, 350 GB of RAM and 7 TB of storage.
One of the main goal of the service was to allow to quickly develop and make available or more accessible to the community novel Public Galaxy based services e.g.:

- **Galaxy CorGAT** dedicated to the use of CorGAT (Coronavirus Genome Analysis Tool) a novel, highly effective and user-friendly approach for the functional annotation of SARS-CoV-2 genomes.
- **Galaxy VINYL** to perform variant prioritization analysis using VINYL (Variant prIoritizatioN bY survivaL analysis) a tool suite that integrates an innovative method for this type of analysis.



Galaxy CorGAT


Galaxy / VINYL
Welcome to VINYL
See the VINYL manual for an explanation of what you can do with VINYL
See this Tutorial for a guided analysis of a reference dataset
VINYL: Variant prIoritizatioN
bY survivaL analysis

Galaxy VINYL

Laniakea@ReCaS Galaxy Community Meeting 2021 Poster

Conclusions

These new Lanikaea software improvements will provide advantages both for users and for system maintainers and developers. The former will have access to a system rich of new applications and frequently updated together with a rapid and error prone instance deployment, while the latter will count on a more stable and efficient system with a reduced maintenance effort.
These updates are currently under deployment, subsequently, after a rigid testing phase, will be ported on the production system Laniakea@ReCaS, to make available to current Laniakea@ReCaS users.

Co-authored publications

- M.Tangaro et al: **Laniakea: an open solution to provide Galaxy "on-demand" instances over heterogeneous cloud infrastructures**, GigaScience, Volume 9, Issue 4, April 2020, giaa033, https://doi.org/10.1093/gigascience/giaa033
- M. Chiara et al **CorGAT: a tool for the functional annotation of SARS-CoV-2 genomes**, Bioinformatics, Volume 36, Issue 22-23, 1 December 2020, Pages 5522–5523, https://doi.org/10.1093/bioinformatics/btaa1047
- M. Chiara et al , **VINYL: Variant prIoritizatioN bY survivaL analysis**, Bioinformatics, Volume 36, Issue 24, 15 December 2020, Pages 5590–5599, https://doi.org/10.1093/bioinformatics/btaa1067
- M. Tangaro, P. Mandreoli et al **Laniakea@ReCaS: exploring the potential of customisable Galaxy on-demand instances as a cloud-based service ,** in press on BMC Bioinformatics

Contacts

- **Pietro Mandreoli** pietro.mandreoli@unimi.it
- **Marco Antonio Tangaro** ma.tangaro@ibiom.cnr.it
- **Federico Zambelli** (ELIXIR-ITA technical coordinator) federico.zambelli@unimi.it