Building a Distributed Computing Network for Galaxy, Application to Genome Annotation

Eva Mercier, Romane Libouban, Thomas Chaussepied, Anthony Bretaudeau

EuroScienceGateway

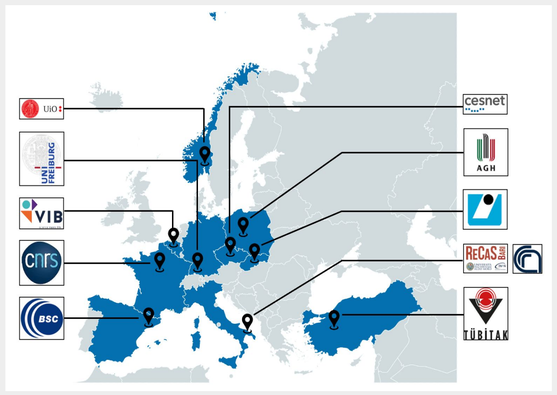
Building a Distributed Computing Network for Galaxy, Application to Genome Annotation

Eva MERCIER[1], Romane LIBOUBAN[1] , Thomas CHAUSSEPIED [1,2] and Anthony BRETAUDEAU[1]

[1] GenOuest, University of Rennes, INRIA, CNRS, IRISA, Rennes, France, [2] CNRS, Institut Français de Bioinformatique, IFB-Core, UAR 3601, 91000, Evry, France

EuroScienceGateway Project Context

Facilitating access to computing and storage infrastructures across Europe in line with the needs of researchers in multiple scientific fields (Biodiversity, Astronomy, Climate and Material Science)



The **GenOuest** platform, in close collaboration with the **Institut Français de Bioinformatique** (IFB), is developing the national **UseGalaxy.fr** server by implementing technologies that enable **interoperability** between the infrastructures available throughout **Europe.** An ESG **goal** is to enhance maturity of **6 national Galaxy servers**, including **UseGalaxy.fr** (TRL-9).

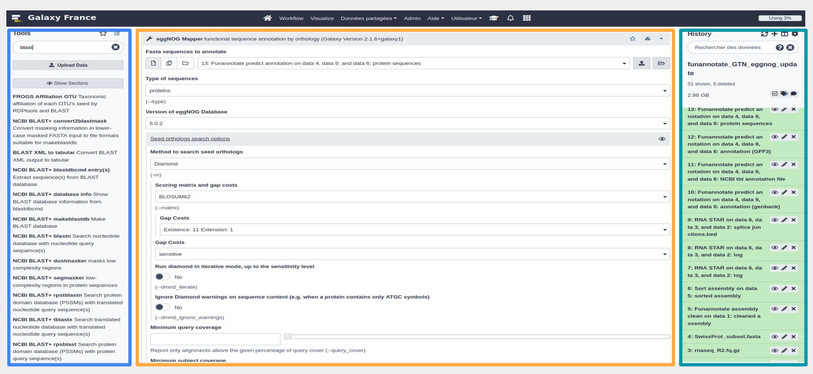
galaxyproject.org/projects/esg

Galaxy: an accessible web portal for FAIR data analysis

useGalaxy.fr

3k active users

4.5M jobs



Common set of tools and workflows, connection to Pulsar endpoints, shared reference data and Apptainer images (CVMFS) , federated login, ...

Local scheduling system

useGalaxy.*

Local scheduling system

useGalaxy.eu

RabbitMQ

RabbitMQ

Message Queuing System

CernVM File System (CVMFS) distributed file system for sharing:

- Reference Data (genomes and indices)
- Tool containers (Biocontainers)

Remote scheduling system

PULSAR

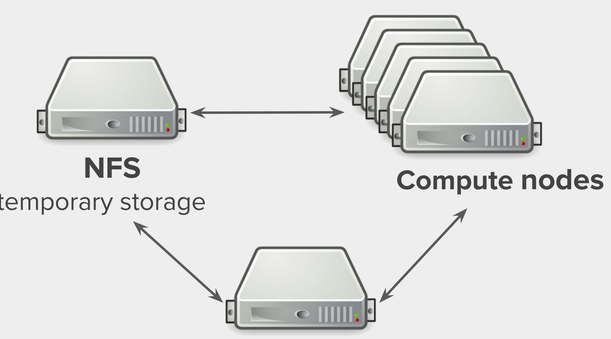
Virtual Galaxy Compute Nodes boot from VGCN images

Standard image including everything required components to run Galaxy jobs

pulsar, docker, singularity, autofs, CVMFS

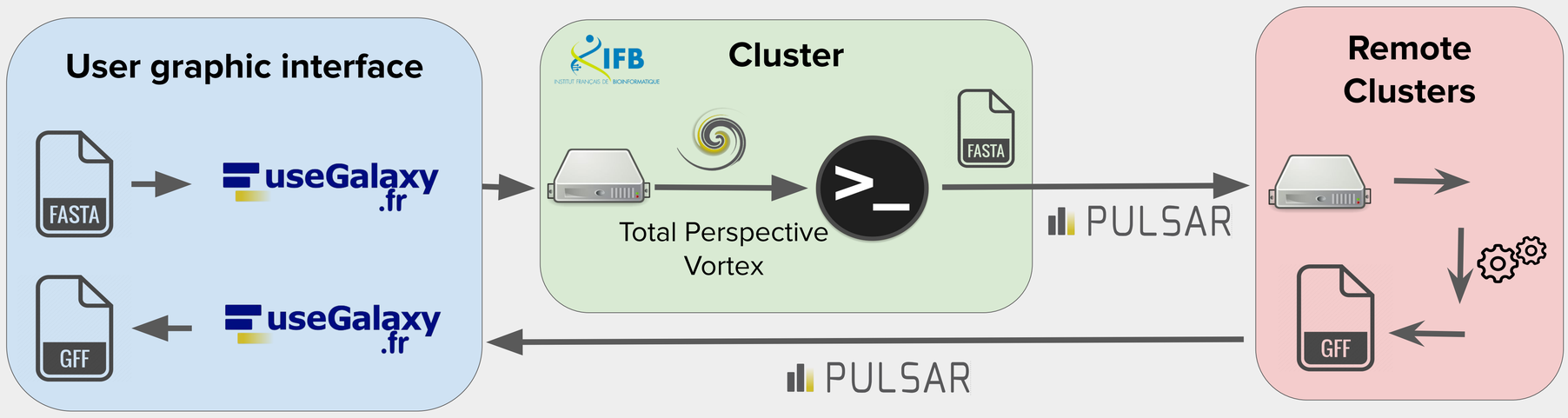
github.com/usegalaxy-eu/vgcn

HTCondor



Job execution with Pulsar

The task is launched from the graphical user interface, then the data is transferred to the IFB cluster. The destination of a pulsar can be configured within the Total Perspective Vortex plugin configuration. The results are sent to the graphical user interface.



Ongoing developments from EuroScienceGateway

Allow users to integrate available external computing and storage resources on their Galaxy server:

- Bring Your Own Compute (BYOC): plug-in personal Pulsar endpoint
- Bring Your Own Storage (BYOS): plug-in remote storage (read and write)
- Manual selection of job destinations

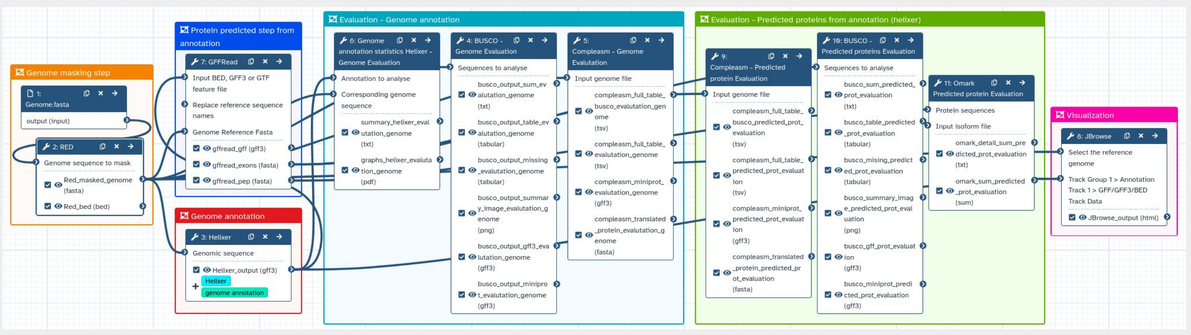
Meta job scheduler to optimise the distribution of jobs, based on:

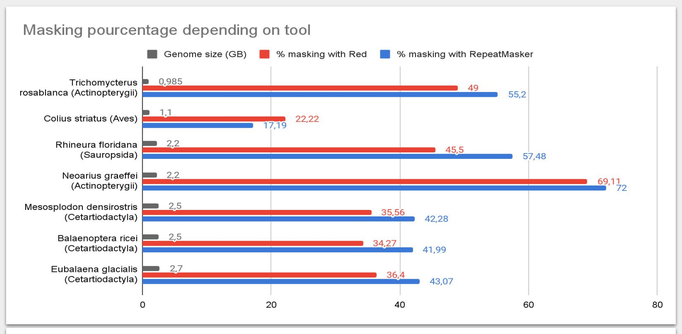
- Data locality (compute near data)
- Computing resources availability (GPU, CPU, memory)
- User preferences / quotas (computing costs)
- Carbon footprint (PUE, energy mix data, deferred execution)

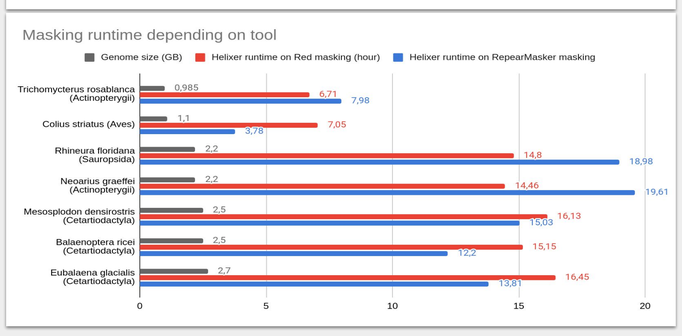
Biodiversity use case: genome annotation

Large scale genome sequencing project
Earth BioGenome Project (EBP) : VGP, ERGA, ATLASea, ...
Need for standard genome annotation procedures

- New Galaxy tools (Braker3, Helixer, Compleasm, OMArk, ...)
- FAIR Workflows (workflowhub.eu/collections/13)
- Online training materials (training.galaxyproject.org/training-material)







Project 2023-2025

Funded by the European Union

EuroScienceGateway was funded by the European Union programme Horizon Europe (**HORIZON-INFRA-2021-EOSC-01-04**) under grant agreement number 101057388 and by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee grant number 10038963.