



**HAL**  
open science

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

Eva Mercier, Romane Libouban, Thomas Chaussepied, Anthony Bretaudeau

► **To cite this version:**

Eva Mercier, Romane Libouban, Thomas Chaussepied, Anthony Bretaudeau. Building a Distributed Computing Network for Galaxy, Application to Genome Annotation. JOBIM 2024 - Journées Ouvertes en Biologie, Informatique et Mathématiques, Jun 2024, Toulouse, France. 2024. hal-04694519

**HAL Id: hal-04694519**

**<https://hal.science/hal-04694519v1>**

Submitted on 11 Sep 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

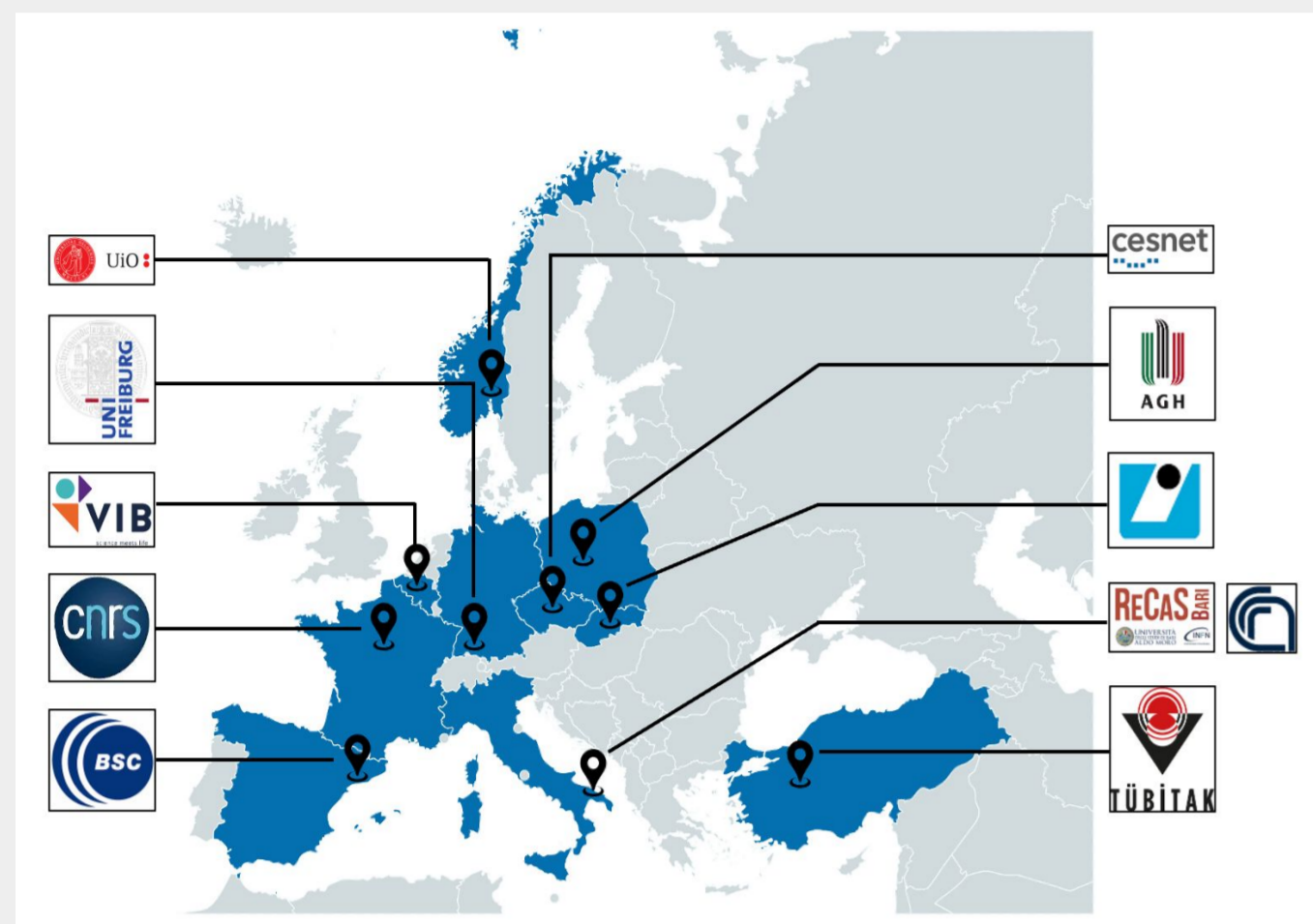
## Building a Distributed Computing Network for Galaxy, Application to Genome Annotation

Eva MERCIER<sup>1</sup>, Romane LIBOUBAN<sup>1</sup>, Thomas CHAUSSEPIED<sup>1,2</sup> and Anthony BRETAUDEAU<sup>1</sup>

<sup>1</sup>GenOuest, University of Rennes, INRIA, CNRS, IRISA, Rennes, France, <sup>2</sup> 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](http://galaxyproject.org/projects/esg)

### Galaxy: an accessible web portal for FAIR data analysis

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

### Local scheduling system



### Local scheduling system



Message Queuing System

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

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

### Remote scheduling system



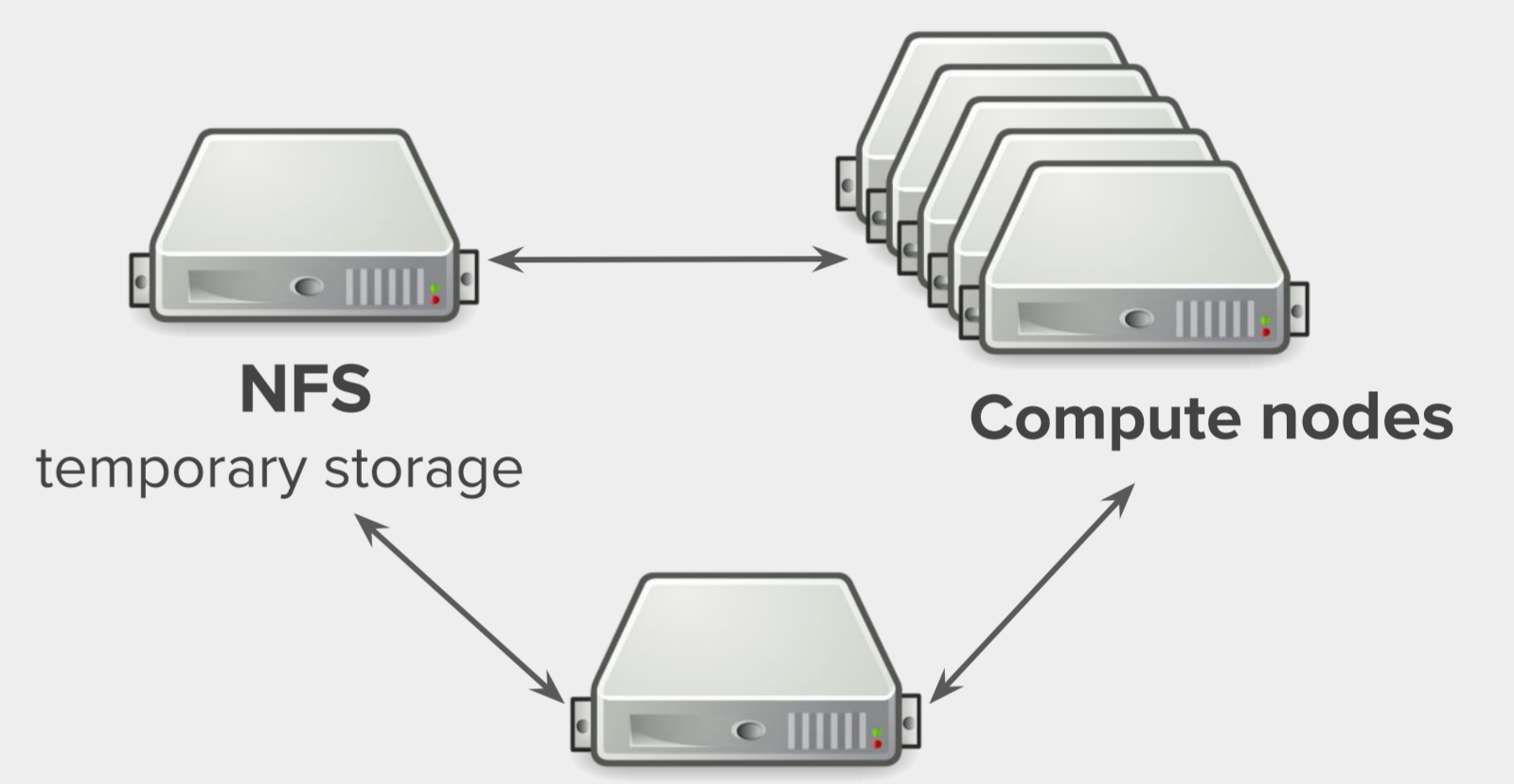
#### Virtual Galaxy Compute Nodes

boot from VGCN images

Standard image including everything required components to run Galaxy jobs

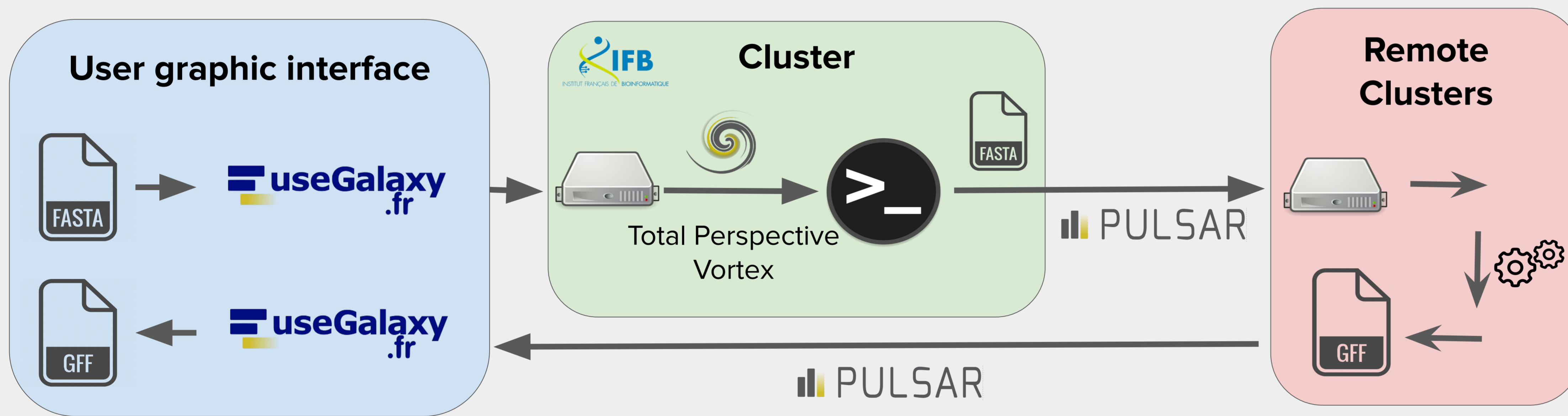
pulsar, docker, singularity, autofs, CVMFS

[github.com/usegalaxy-eu/vgcn](https://github.com/usegalaxy-eu/vgcn)



### 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

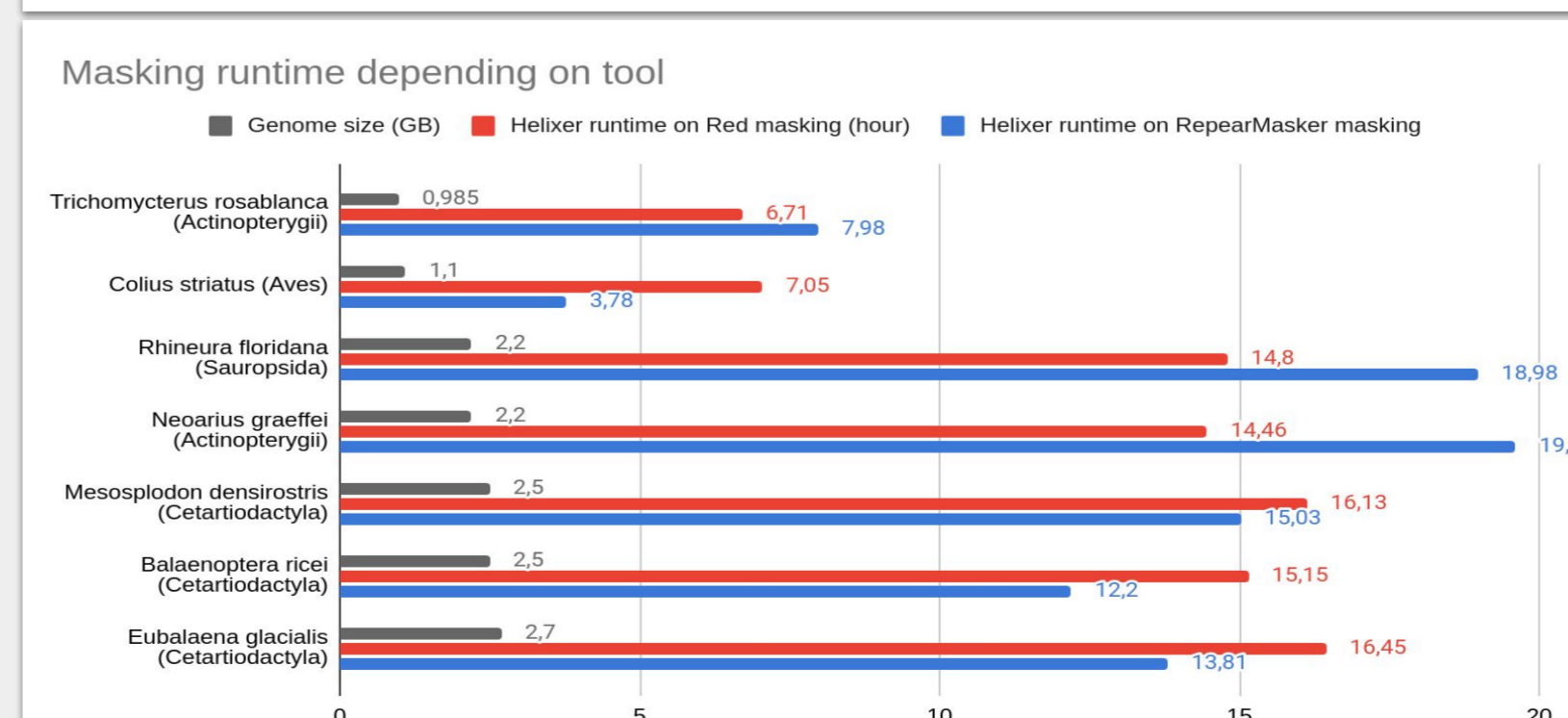
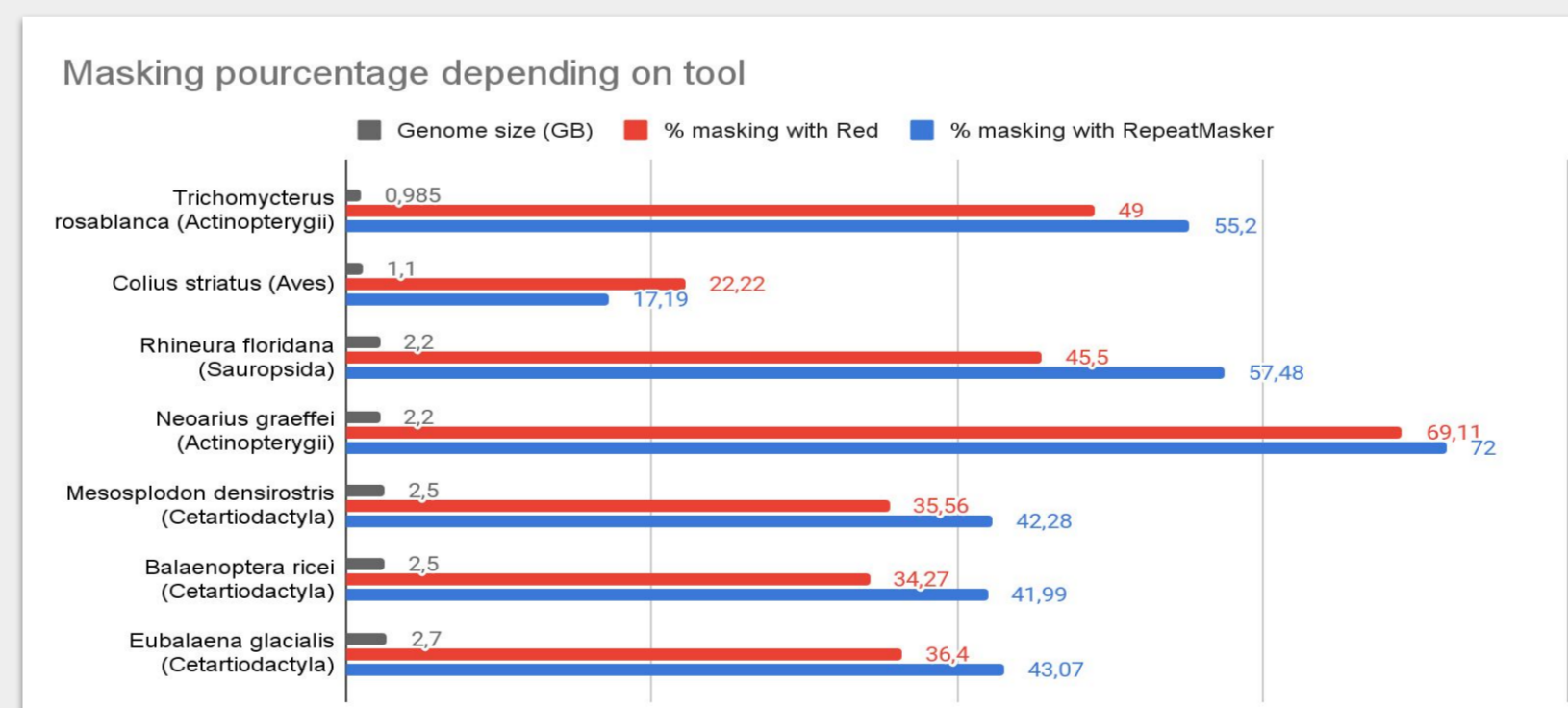
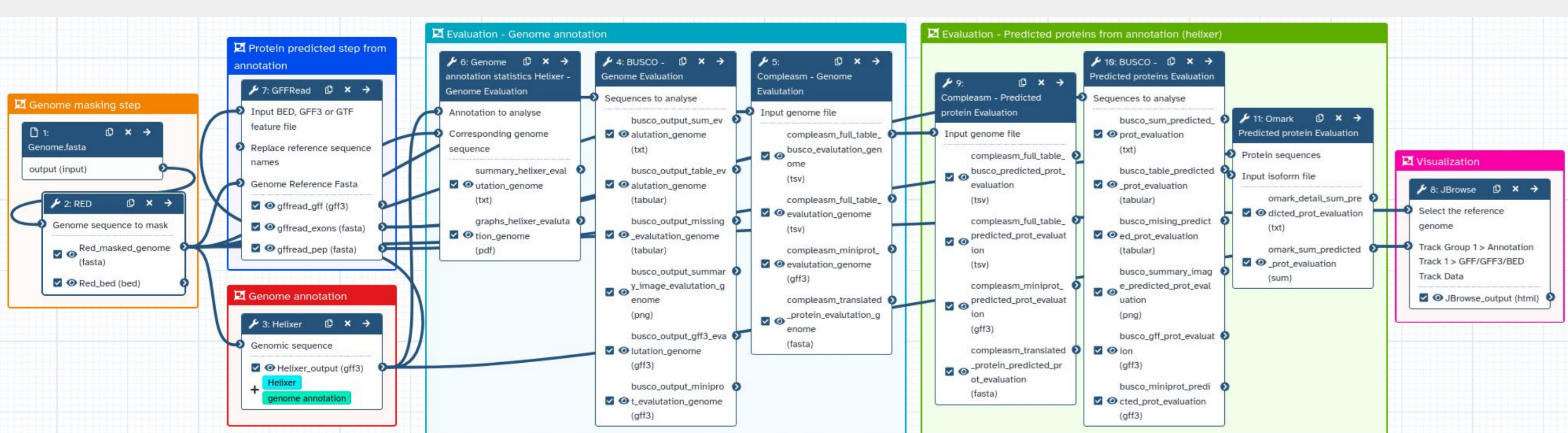
Allow users to integrate available external computing and storage resources on their Galaxy server:

- 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

### Biodiversity use case: genome annotation

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

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



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)