



# Open scientific computing with Galaxy



WLCG Grid Deployment Board Presentation  
February 9, 2022

[www.elixir-europe.org](http://www.elixir-europe.org)



The image shows a world map with three Galaxy logos placed over different continents. Each logo consists of a stylized 'G' icon followed by the word 'Galaxy' and a region name below it. Orange double-headed arrows connect the logos: one between North America and Europe, one between Europe and Australia, and one between North America and Australia.

**Galaxy**  
Main

**Galaxy**  
EUROPE

**Galaxy**  
AUSTRALIA

# Objectives

- We are facing similar problems than you do!
- Very large set of data of small'ish data vs. small set of large data
- Underlying hardware and software solutions are the same
- Can we join forces? (**on distributed compute!**)



# It will only grow...

**Table 1. Four domains of Big Data in 2025.** In each of the four domains, the projected annual storage and computing needs are presented across the data lifecycle.

<u>Data Phase</u>	<u>Astronomy</u>	<u>Twitter</u>	<u>YouTube</u>	<u>Genomics</u>
<b>Acquisition</b>	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
<b>Storage</b>	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
<b>Analysis</b>	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
<b>Distribution</b>	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001



# All data-rich disciplines have issues with ...

**Accessibility:** Making use of large-scale data requires complex computational resources and methods. Can all researchers access these approaches? How can we make these methods available to *everyone*

**Transparency:** Is it possible to communicate analyses and results in ways that are both easy to understand and provide all of the essential details

**Reproducibility:** Can analyses be precisely reproduced, to facilitate rigorous validation and peer review, and ease reuse?



A generic, open science research  
environment

-

agnostic w.r.t. data or scientific domain

# Galaxy is a collection of mature components

- Integration and maintenance of tools and workflows
- Creating and distributing containerized tools
- SDK
- Comprehensive API
- Distributed computing middleware
- Abstraction layers for storage, vis, AAI ...
- World-wide training hub



PLANEMO

PULSAR



# Why use Galaxy for scientific computing?

## The brief explanation

So scientists can choose replicable tools and automated workflows for their analyses...

*...in addition to a command line if they need it.*

RNA STAR Gapped-read mapper for RNA-seq data (Galaxy Version 2.7.8a+galaxy0)

Single-end or paired-end reads

Single-end

RNA-Seq FASTQ/FASTA file

899: all\_consensus.fasta.gz (as fasta)

Custom or built-in reference genome

Use a built-in index

Built-ins were indexed using default options

Reference genome with or without an annotation

use genome reference without builtin gene-model

Select the '...' with builtin gene-model' option to select from the list of available indexes with splice junction information. Select the '...' without builtin gene-model' option to select from the list of available indexes without annotated splice junctions, and, optionally, provide your own gene model file.

Select reference genome

A. mellifera genome (apiMel3, Baylor HGSC Amel\_3.0)

If your genome of interest is not listed, contact the Galaxy team (--genomeDir)

Gene model (gff3,gtf) file for splice junctions

Nothing selected

Exon junction information for mapping splices (--sjdbGTFfile)

jupyter Untitled Last Checkpoint: a minute ago (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Trusted Python 3

In [ ]: |





# Galaxy: Open Computing for Researchers

- Generic Research Environment
  - Compose platform via AppStore (8k+ tools)
- Easily adapted: Used in many data rich disciplines.
  - Genomics; Images; ML; Climate; Chemoinfo,...
- Free open analysis services or local server
- Integrated training
- >11k citations
- Active, supportive community



# Open, repeatable scientific computing

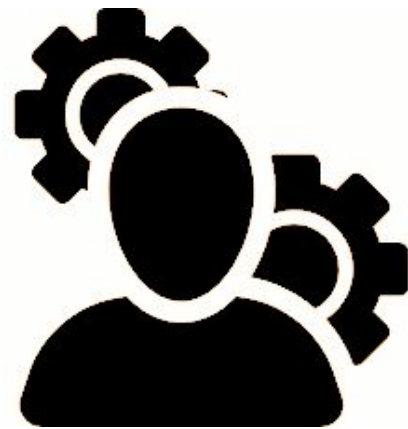
Scrutiny, replication required for scientific trustworthiness

- Open source - framework, utilities and tools
- Shareable jobs, workflows and completed analyses
  - for collaboration or peer review
  - for support - bugs also replicable!
- Capture all provenance to enable full computational replication



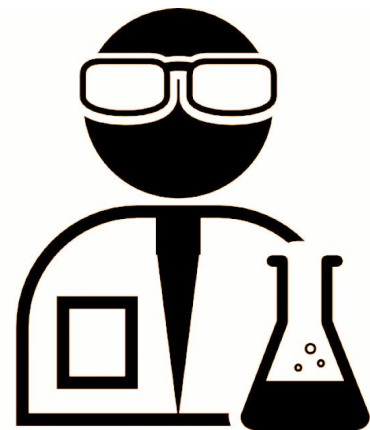
# What Service Providers see

- Efficient use of large hardware allocations
- Global Community providing support, training and advice
- Professional, open software engineering
- Automated tool/dependency management
- Any command line package → shareable tool
- API for local service integration
- *Highly available open computing services for scientists*



# What Scientists see

- On line analysis platform with integrated training
- Deploy locally or use public open analysis services
- An active, supportive community
- Manage own data, workspaces and workflows
  - Web browser GUI. Jupyter notebook option
  - Replicable computational jobs
  - Shareable data, workflows, analyses
- *Easy access to open computing analyses*



# Galaxy relies on CERN components

- CVFMS serves project wide resources
  - 1000's of reference genomes - e.g. mouse
  - Container images for tools
  - Minimises duplication at a *global scale*
- Zenodo/Invenio
  - GTN data and resources
  - Citation harvesting:  
<https://galaxyproject.org/publication-library/>



# CVMFS server distribution

- Stratum 0 servers
- Stratum 1 servers

cvmfs0-psu0

- singularity.galaxyproject.org
- data.galaxyproject.org

cvmfs1-psu0

Penn State

XSEDE, Indiana University

cvmfs1-iu0

cvmfs0-tacc0

- test.galaxyproject.org
- main.galaxyproject.org

cvmfs1-tacc0

XSEDE & CyVerse,  
TACC, Austin

cvmfs1-ufr0.usegalaxy.eu

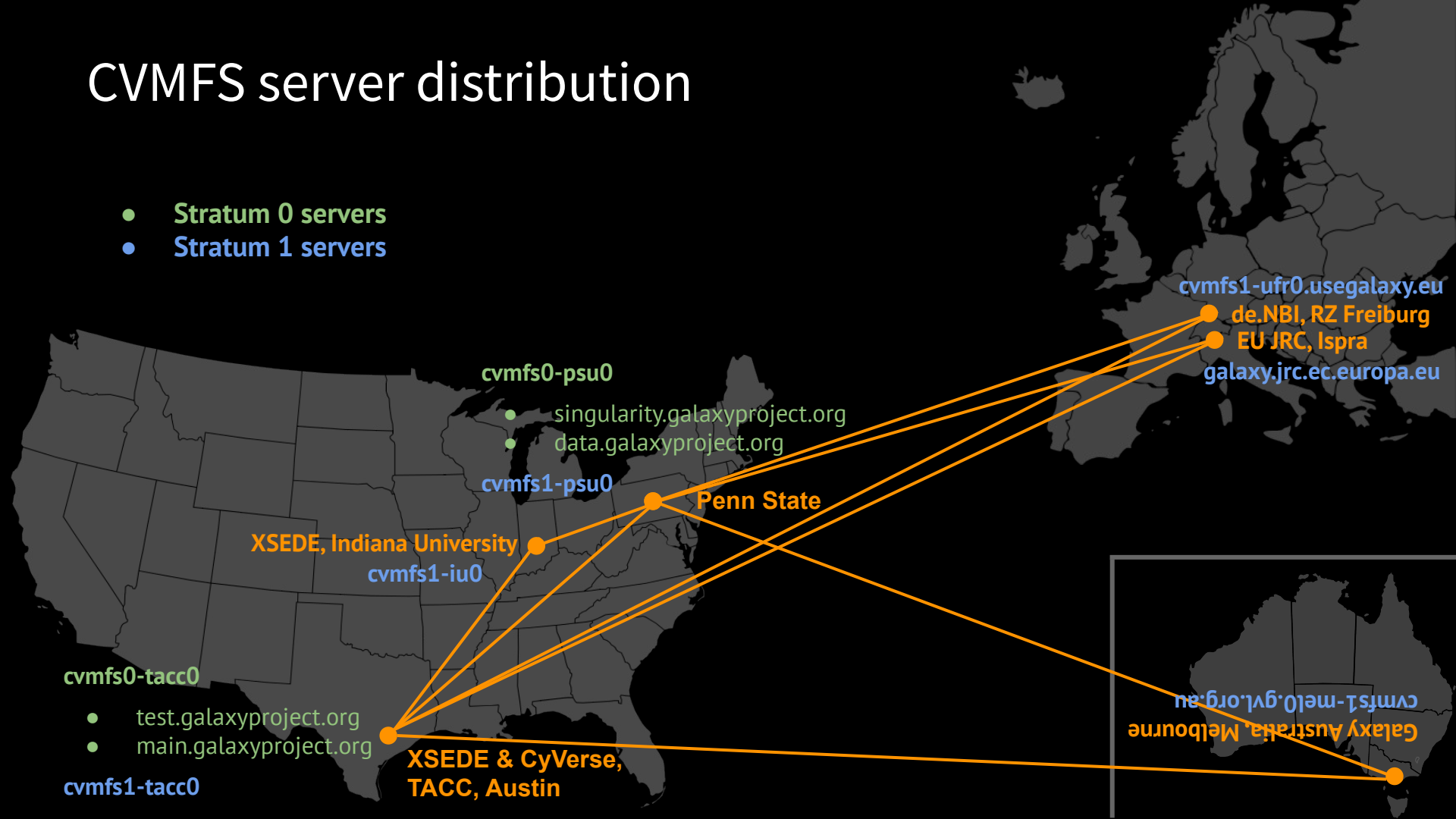
de.NBI, RZ Freiburg

EU JRC, Ispra

galaxy.jrc.ec.europa.eu

cvmfs1-mel0.gvl.org.au

Galaxy Australia, Melbourne

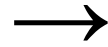


Worldwide  
resources

200 public  
Galaxy servers

8,000 tools available  
in AppStore

IaaS



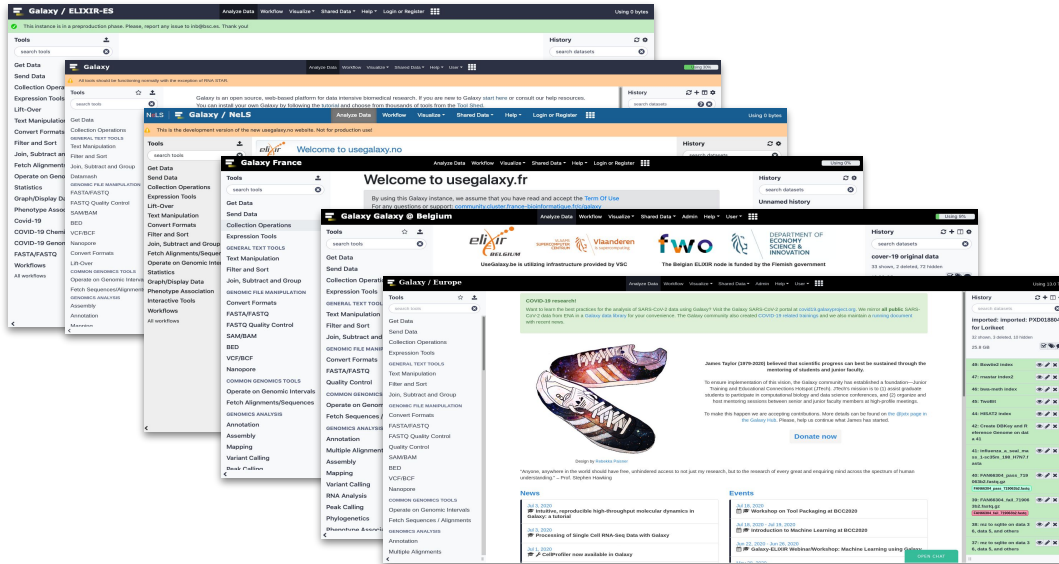
PaaS



SaaS



# Free open analysis services



usegalaxy.org  
usegalaxy.org.au  
usegalaxy.eu  
usegalaxy.fr  
usegalaxy.be  
usegalaxy.ee  
usegalaxy.es  
usegalaxy.it

## 138 platforms for using Galaxy





# Many research scenarios

- 10's of thousands of large datasets | VGP
- 100's of thousands of datasets | COVID-19
- Analysis of protected human data | AnVIL
- ML using large image datasets | Tumor maps in Cancer Research

## Machine Learning

The Machine Learning section contains three sub-panels:

- Fully connected layer:** A diagram showing a layer of nodes with connections between them.
- Confusion matrix:** A square matrix with a diagonal line of green cells, representing classification performance.
- Classifier Decision Function:** A graph showing two overlapping normal distribution curves (one blue, one red) on a probability axis from 0.0 to 1.0.

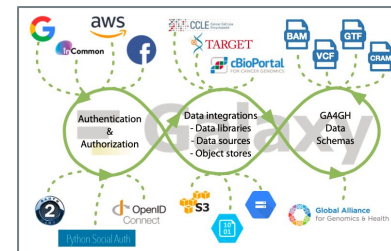
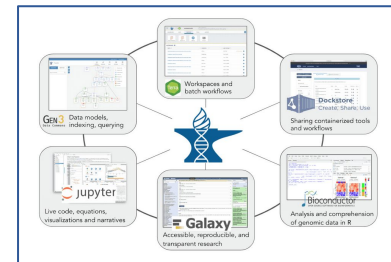


### correspondence

#### Ready-to-use public infrastructure for global SARS-CoV-2 monitoring

The flowchart illustrates the infrastructure for global SARS-CoV-2 monitoring, showing the flow from data collection to analysis and reporting.

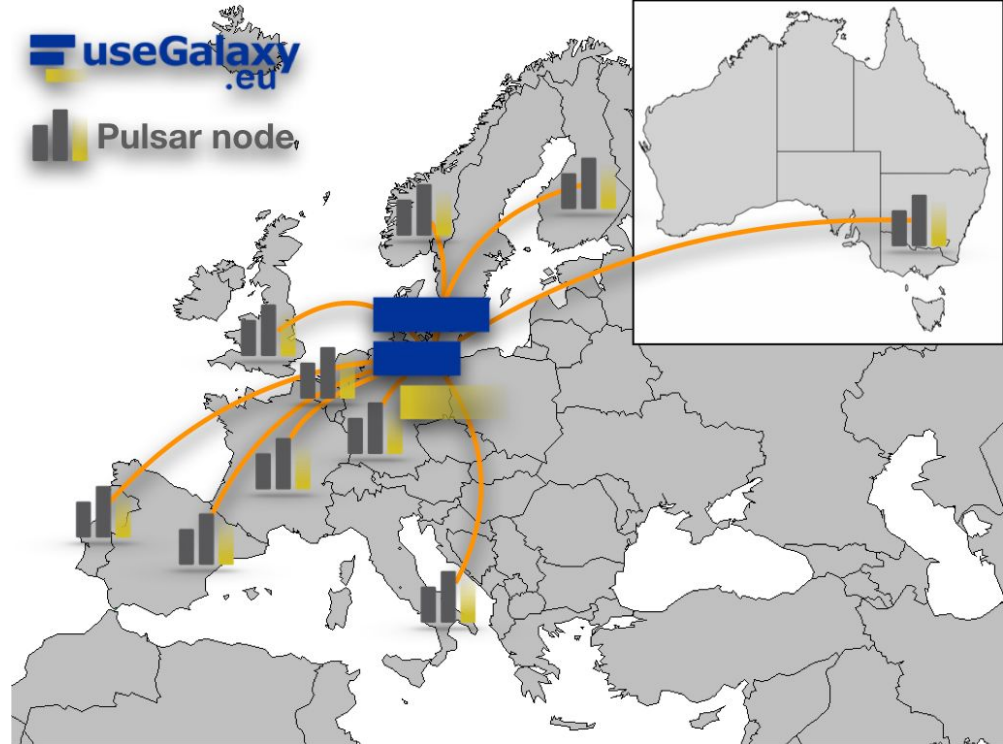
**Fig. 1 | Analysis flow for calling SARS-CoV-2 variants using Galaxy, CNT, Oxford Nanopore Technologies, VCF, variant call format, TSV, tab-separated values, FC, panel vcf, SE, single end, for more information, see <https://www.nature.com/articles/d41586-021-00000-0>**



# The PULSAR Network

- Distributing the computational load of continental Galaxy servers
- Access to diverse hardware e.g. GPU cluster in UK
- E.g. to analyse COVID-19 data

<https://pulsar-network.readthedocs.io/en/latest/project/partners.html>



# Welcome to Galaxy Training!

Collection of tutorials developed and maintained by the worldwide Galaxy community

## Galaxy for Scientists

Topic	Tutorials
<a href="#">Introduction to Galaxy Analyses</a>	9
<a href="#">Assembly</a>	12
<a href="#">Climate</a>	4
<a href="#">Computational chemistry</a>	7
<a href="#">Ecology</a>	6
<a href="#">Epigenetics</a>	7
<a href="#">Genome Annotation</a>	9
<a href="#">Imaging</a>	4
<a href="#">Metabolomics</a>	6
<a href="#">Metagenomics</a>	7
<a href="#">Proteomics</a>	26
<a href="#">Sequence analysis</a>	3
<a href="#">Statistics and machine learning</a>	15
<a href="#">Transcriptomics</a>	32
<a href="#">Variant Analysis</a>	10
<a href="#">Visualisation</a>	2

## Welcome to the GTN!

Find out more about Galaxy Training Network



Video created by Geert Bonamie.

## The latest GTN news



Read about new tutorials, features, events and more!

- 2022-03-14 - Mar 18, 2022**  
GTN Smörgåsbord 2: Tapas Edition
- Dec 14, 2021**  
Support for annotating Funding Agencies
- Dec 1, 2021**  
New Tutorials: PacBio data QC and Genome Assembly, and Genome Annotation with Funannotate
- Dec 1, 2021**  
New FAQs: How does the GTN stay FAIR and Collaborative

*training.galaxyproject.org*  
Very useful place to start



*The Gallantries, Galaxy Training Network &  
Galaxy Community are happy to announce*

# GTN Smörgåsbord 2

## 14-18 March 2022

Save the date!  
[bit.ly/smorgasbord2](https://bit.ly/smorgasbord2)

Join a **free, global**, week-long Galaxy Training event covering everything from RNA-Seq, Single Cell, Proteomics, SARS-CoV-2 *and more!* This year will include **Galaxy Admin Training.**

🐦 @gxytraining @Gallantries\_EU



With the support  
of the  
European Union

**Galaxy administrator training week coming up!**  
On-line expert tutors; self directed learning materials available in the GTN





The image shows a world map with three Galaxy logos placed over different continents. Each logo consists of a stylized 'G' icon followed by the word 'Galaxy' and a region name below it. Orange arrows connect the logos in a clockwise cycle: from Europe to Australia, from Australia to Main, and from Main to Europe.

**Galaxy**  
Main

**Galaxy**  
EUROPE

**Galaxy**  
AUSTRALIA

# Questions?



The contributor community sustaining Galaxy



fin

Align technologies across domains where possible

Future of federated analysis

Future of distributed storage

Sensitive data

# Abstraction layers

data provenance storage  
(SQL-DB)

Data Types and metadata

Compute  
(HPC, Cloud-Computing,  
GPGPU ...)

Storage (Posix, ObjectStore,  
iRODS ...)

User Interfaces  
(different communities)

Data exports  
(bag-it, RO-crate ...)

Tools (8000)

Reference data

Visualisations

Tool installation  
(modules, Docker, Singularity,  
Conda, ...)

Data imports

Authentication & Authorization

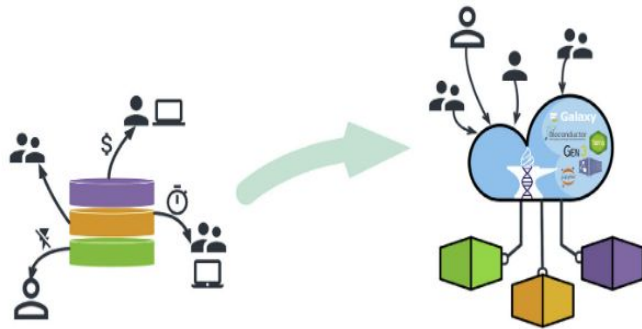




# US secured deployment

- Galaxy used for protected human data analyses in AnVIL
- Inverted model: Compute → Data saves downloading
- National Human Genome Research Institute (US NIH)
- US FISMA certified for this secure use

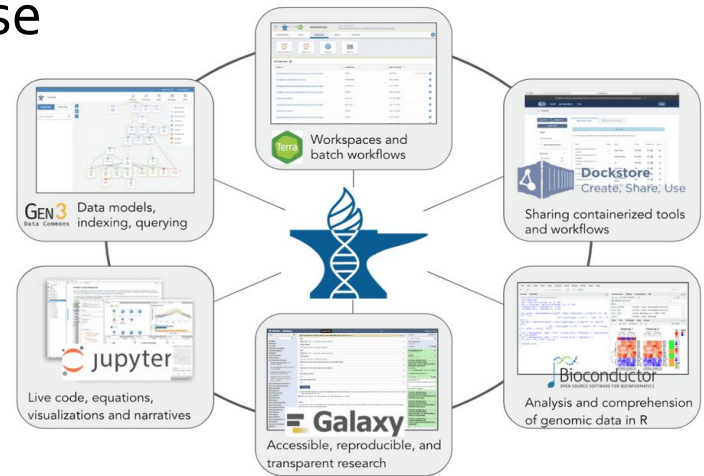
Cell Genomics  
Perspective



CellPress  
OPEN ACCESS

**Figure 1. Inverting the model for data sharing**

(Left) In the traditional model, project data (shown in purple, orange, and green) are copied to multiple sites where they are accessed by users on institutional computing clusters. Under this model, each institution must establish its own data center, and collaboration is achieved primarily through copying files between data centers. (Right) In the inverted model, users connect to a cloud-enabled resource such as the AnVIL to remotely access and analyze the data without copying. In this model, users virtually access a unified data center, allowing for deeper collaboration and sharing of the results.

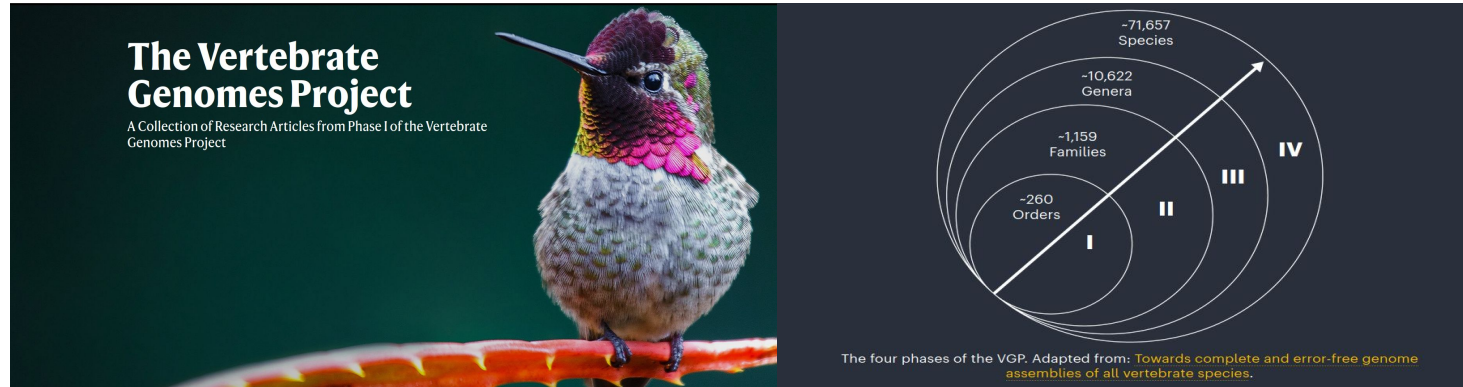


eliXir

# Open computing: Vertebrate Genomes Project

- Plan to sequence and assemble 70,000 high fidelity genomes
- ~1TB input data, 5000 core hours for a typical genome assembly
- Commercial assembly compute cost ~ sequencing cost
- Collaboration: Tools wrapped; Workflows in Galaxy
- Can now use AnVIL Galaxy described above

nature



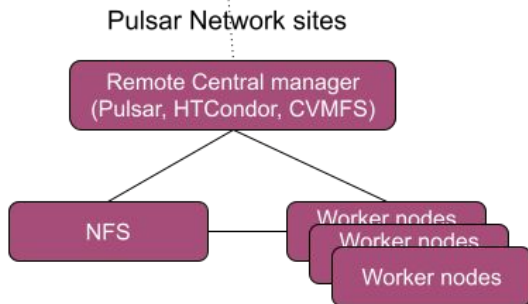
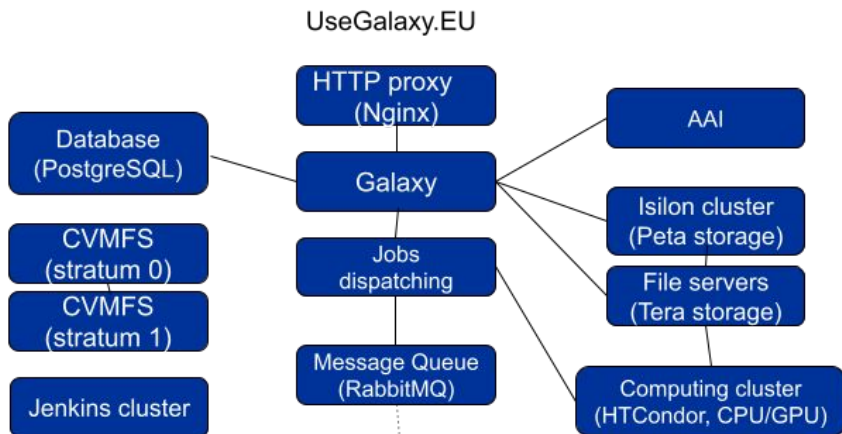


Boxes OS.: Centos8

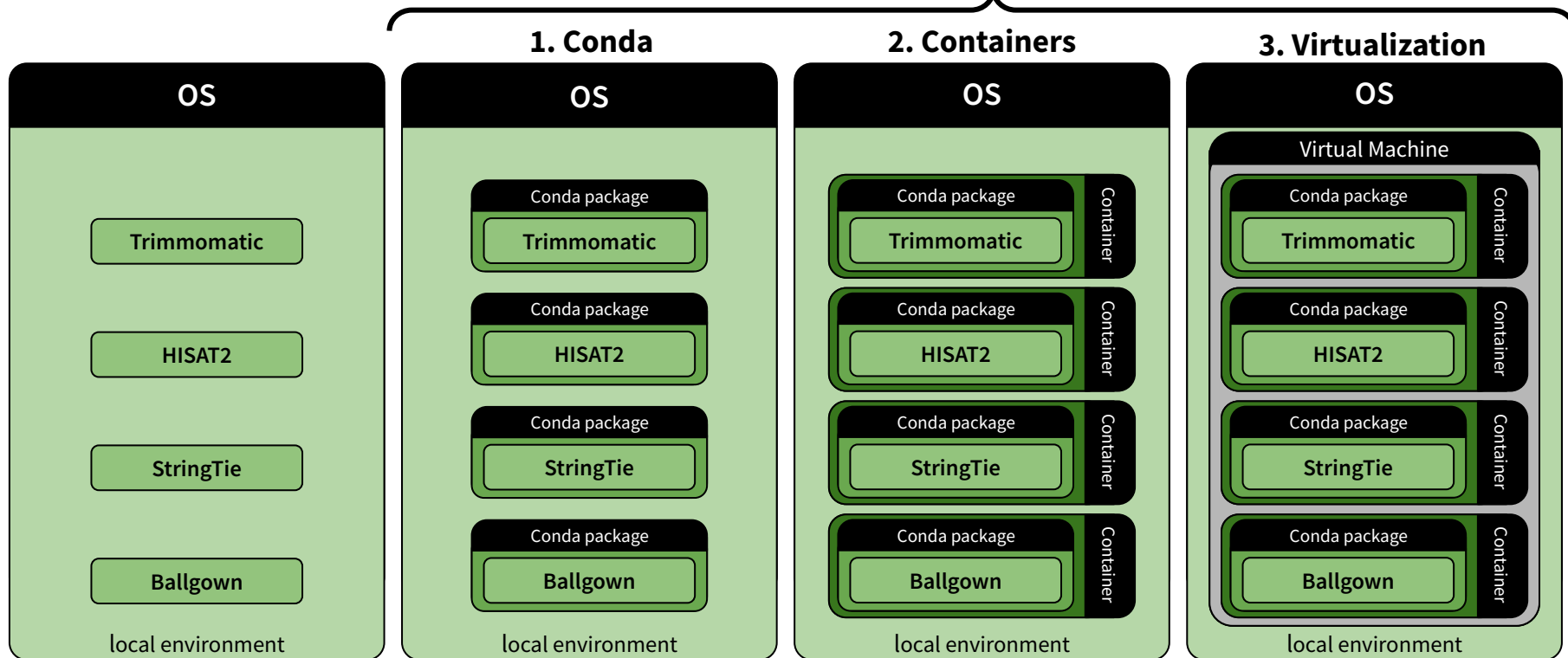
Galaxy: 40 cores, 256GB, 240GB SSD

Database: 40 cores, 256GB, 2TB Raid1 SSD

Jenkins: 40 cores, 128GB, 240GB SSD



# Reproducibility stack



Least reproducible /  
secure

Most reproducible /  
secure

```
> git clone hisat2  
> make  
> sudo make install  
> hisat2 --version
```

```
> conda install hisat2  
> hisat2 --version
```

```
> docker run --rm  
quay.io/biocontainers/  
hisat2 --version
```



# Galaxy as an enabler in Australia

- A highly accessible platform for our 30,000 life sciences researchers (+200,000 students)
  - opened the door for many non-technical users
- Enabled aggregated national compute power
  - bypassing life-sciences-unfriendly allocation mechanisms
  - accessible portal to resource-hungry applications - 1TB nodes
- Standardised vehicle for real-time distribution of global tools and workflows - effectively a global marketplace
- Training as outreach mechanism locally & regionally

