Open scientific computing with Galaxy

WLCG Grid Deployment Board Presentation
February 9, 2022

www.elixir-europe.org


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.

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	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
Distribution	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?

Galaxy

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

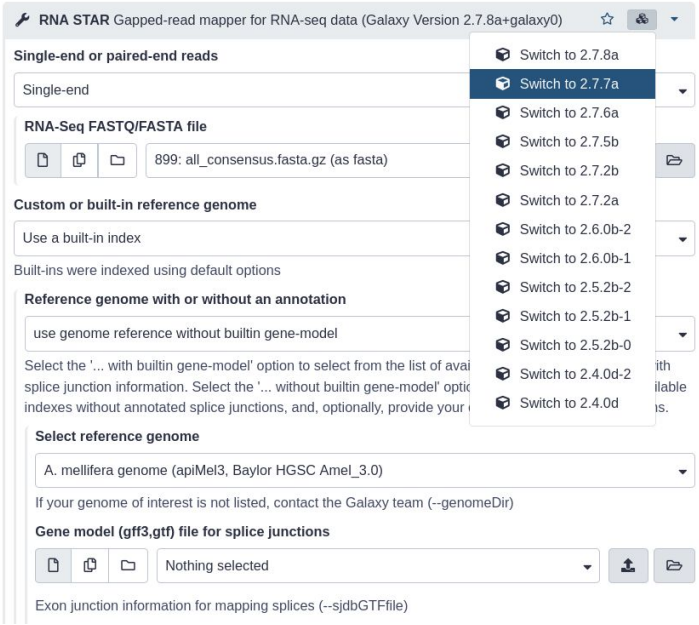
PULSAR

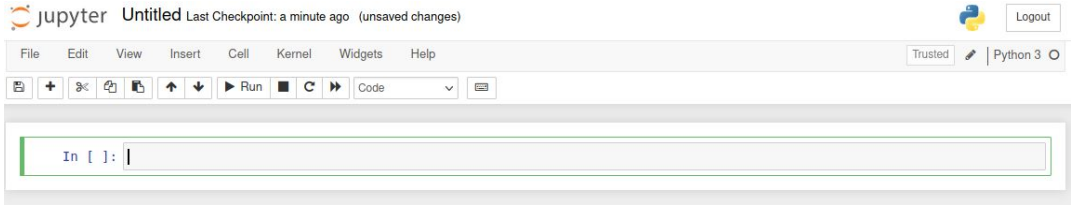
Why use Galaxy for scientific computing?

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.





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

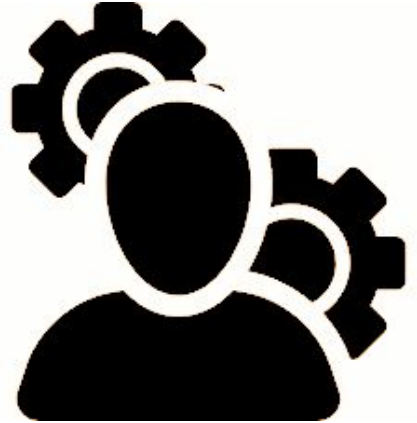
Open, repeatable scientific computing

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

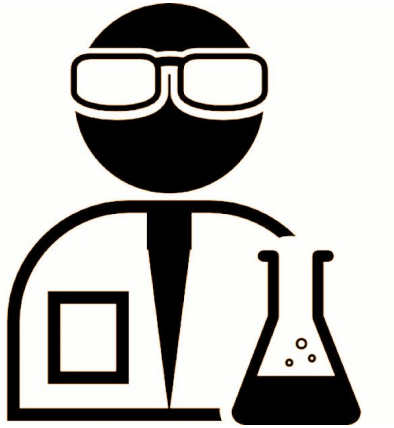
What Service Providers see

- 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

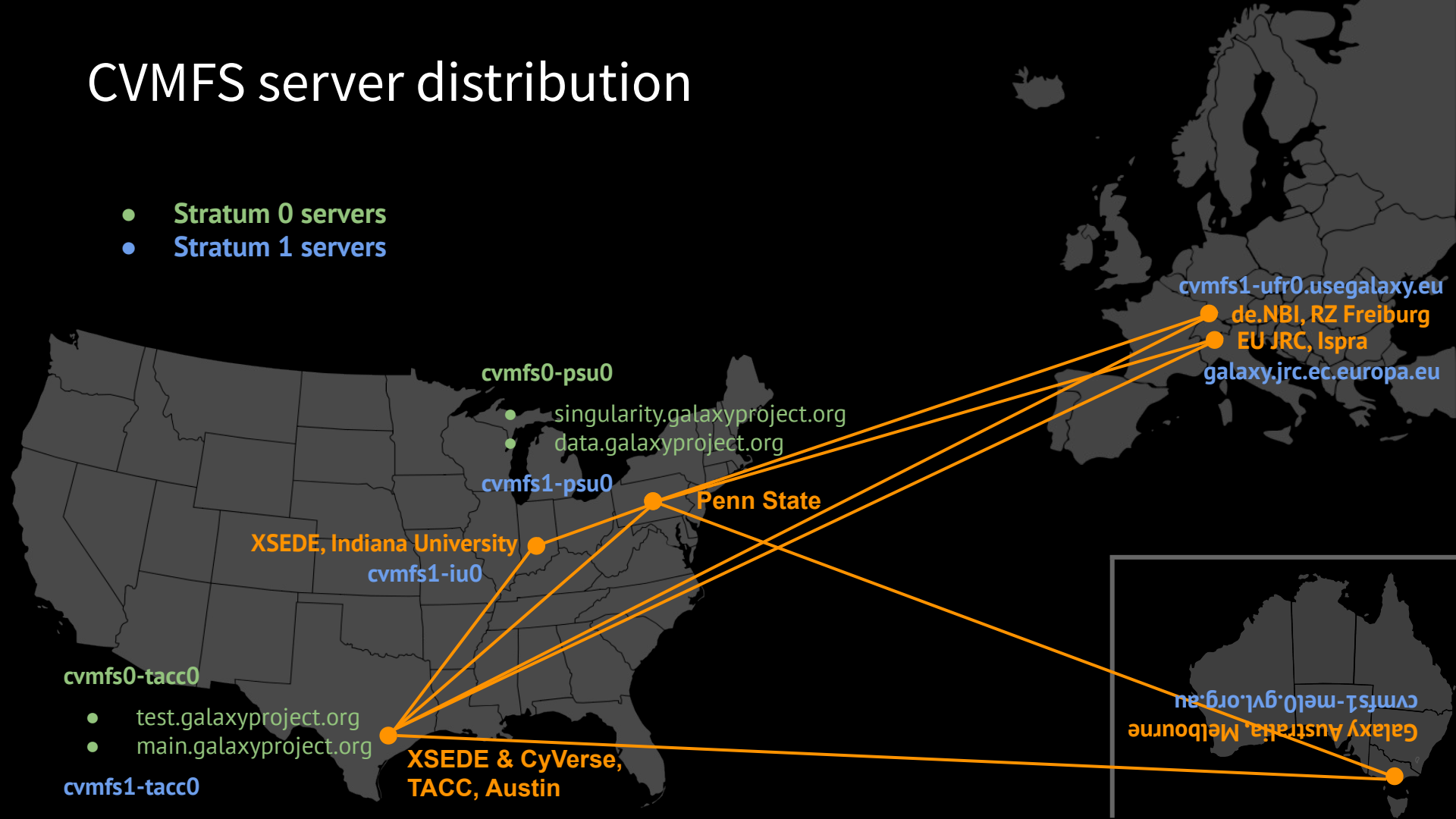
- On line analysis platform with integrated training
- Deploy locally or use public open analysis services
- An active, supportive community
- Manage own data, workspaces and workflows
  - Web browser GUI. Jupyter notebook option
  - Replicable computational jobs
  - Shareable data, workflows, analyses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



Worldwide resources

200 public Galaxy servers

8,000 tools available in AppStore

IaaS → PaaS → SaaS

EUROPEAN OPEN SCIENCE CLOUD

EGI

TACC

BIOPLATFORMS AUSTRALIA

XSEDE
Extreme Science and Engineering Discovery Environment

AnVIL

Galaxy

PULSAR

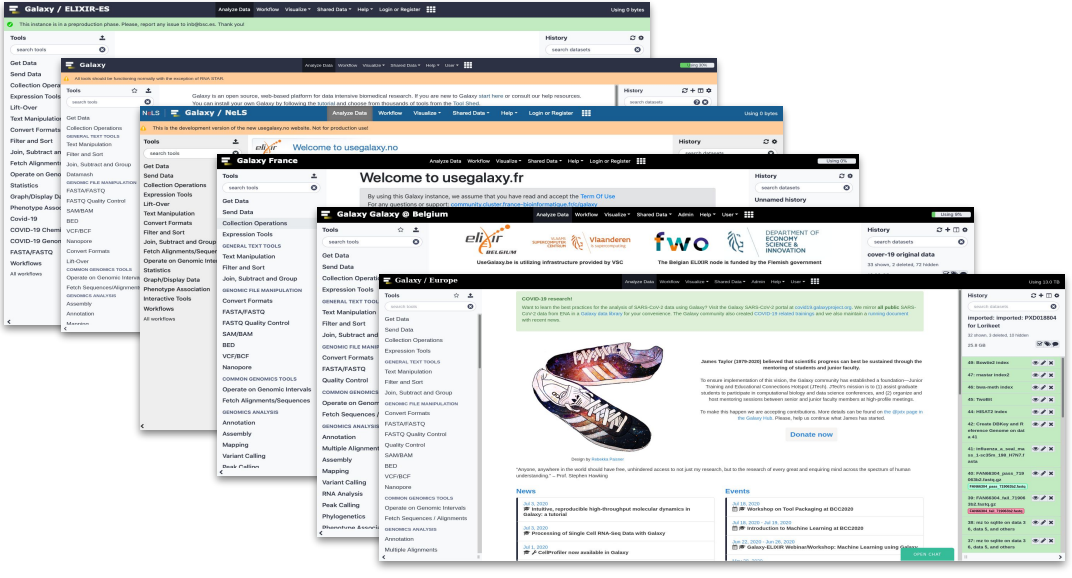
BIOCONDA

CONDA-FORGE

docker

elixir

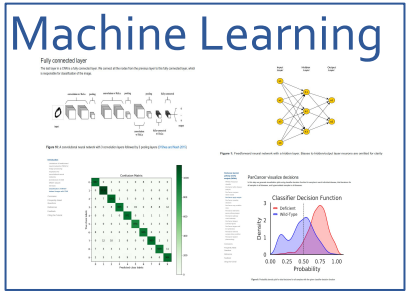
Free open analysis services

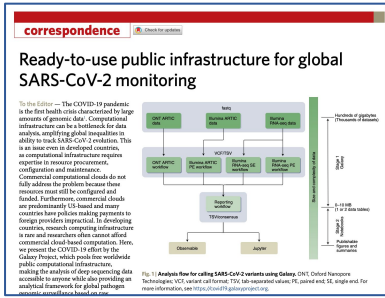


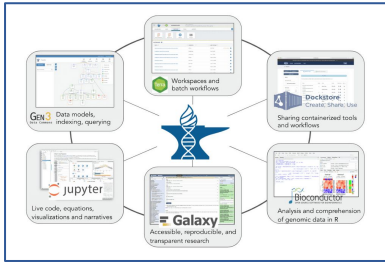
usegalaxy.org
usegalaxy.org.au
usegalaxy.eu
usegalaxy.fr
usegalaxy.be
usegalaxy.ee
usegalaxy.es
usegalaxy.it

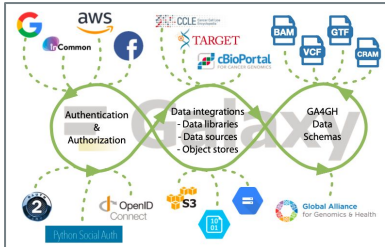
Many research scenarios

- 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





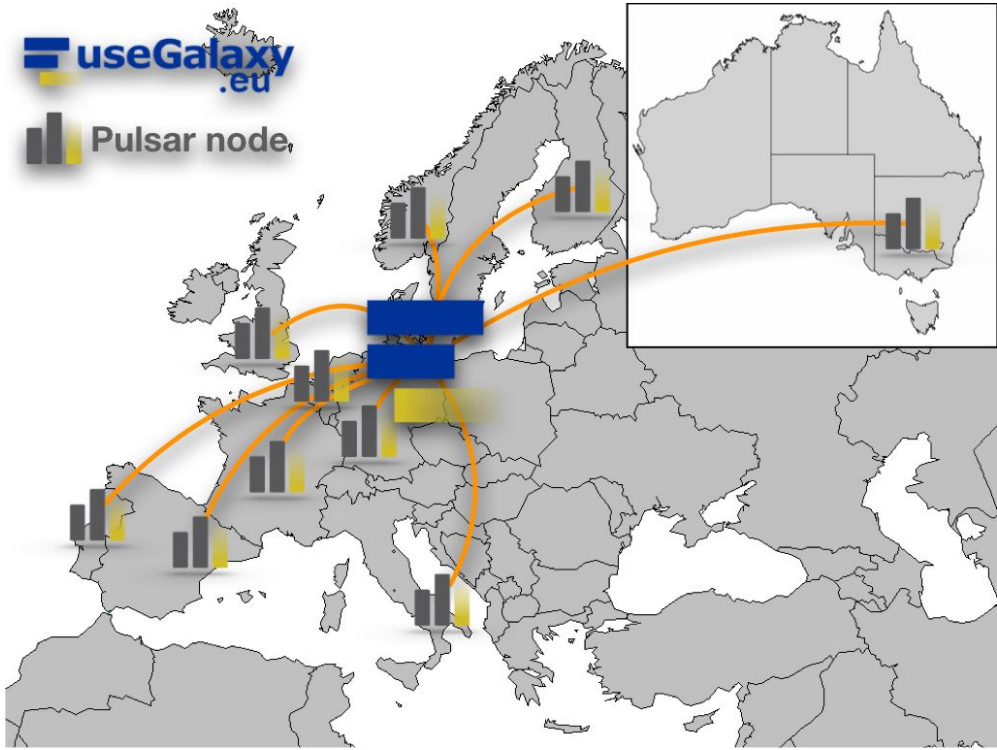


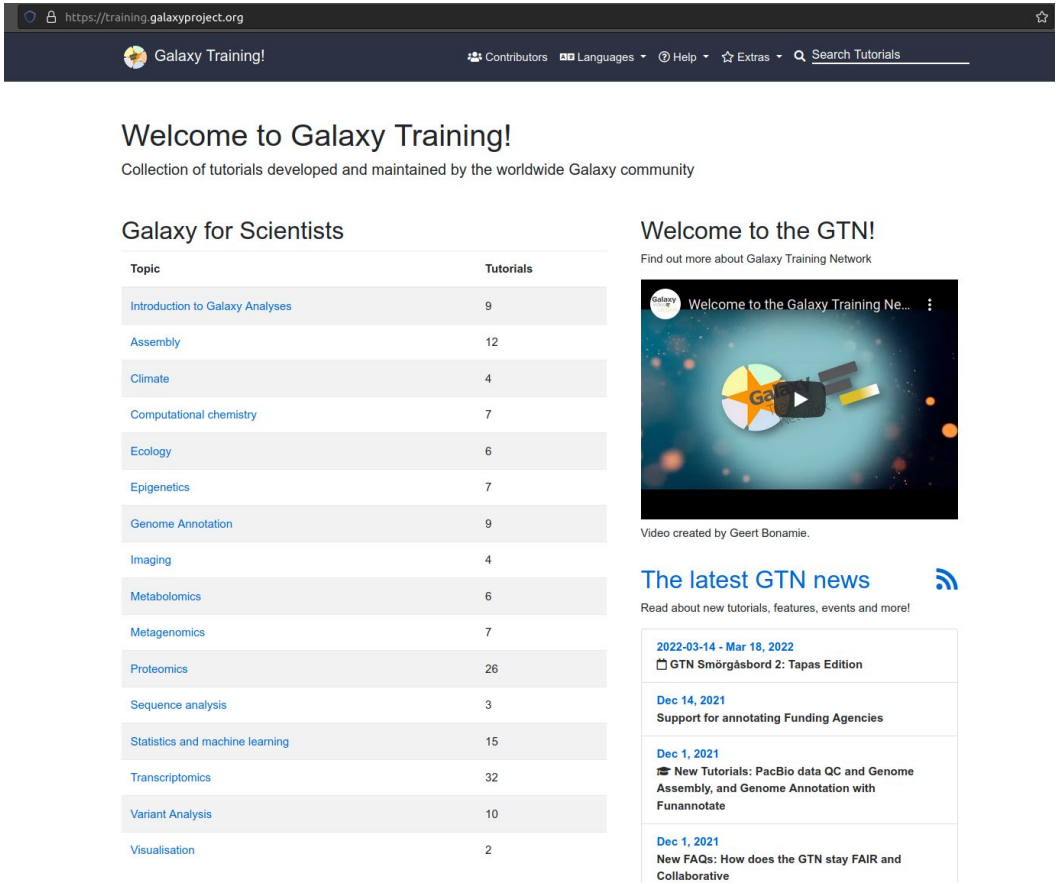


The PULSAR Network

- 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





training.galaxyproject.org

Very useful place to start

Galaxy administrator training week coming up!

On-line expert tutors; self directed learning materials available in the GTN


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

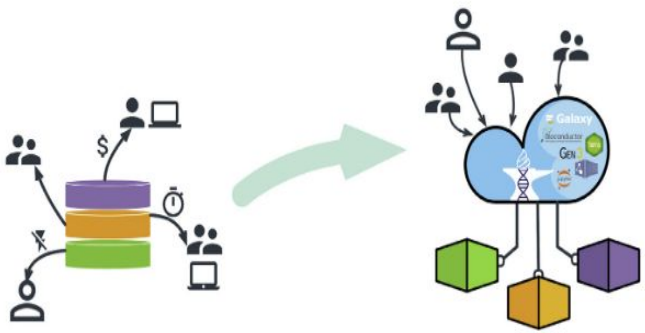
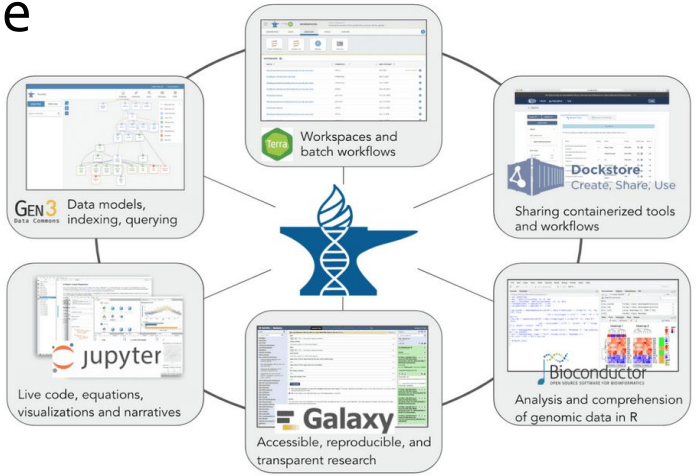


Figure 1. Inverting the model for data sharing

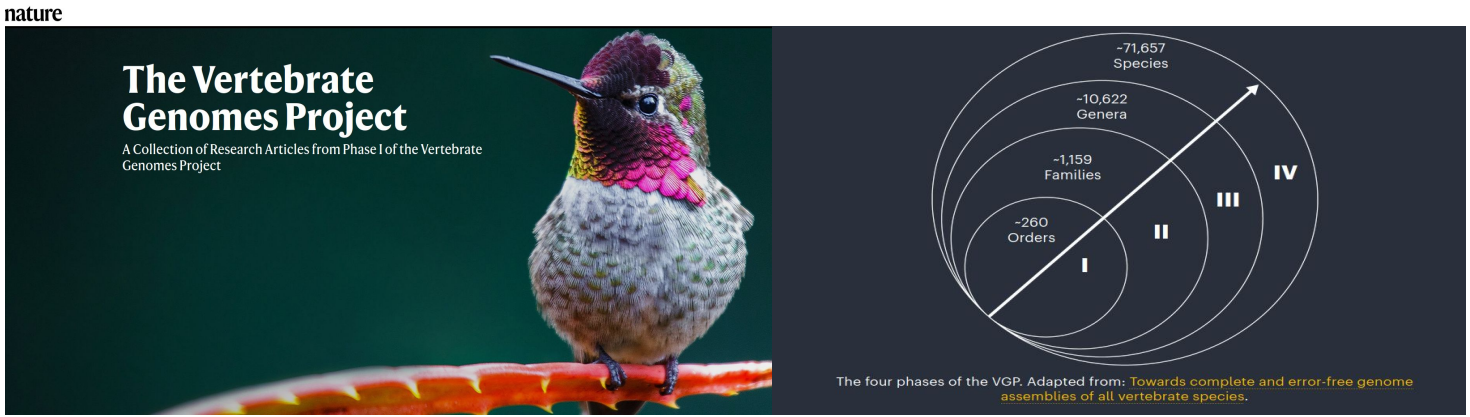
(Left) In the traditional model, project data (shown in purple, orange, and green) are copied to multiple sites where they are accessed by users on institutional computing clusters. Under this model, each institution must establish its own data center, and collaboration is achieved primarily through 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.



Schatz, Philippakis *et al.* (2022) Cell Genomics. https://doi.org/10.1016/j.xgen.2021.100085

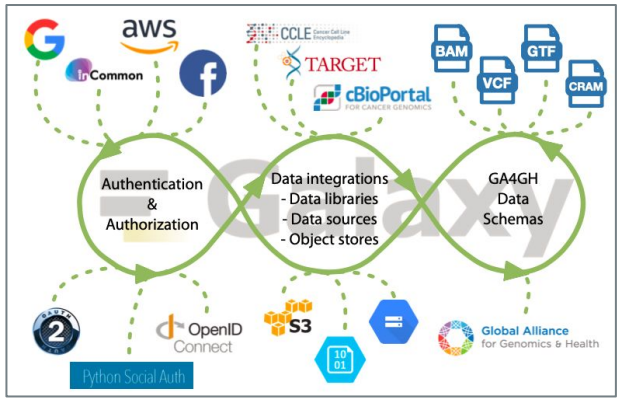
Open computing: Vertebrate Genomes Project

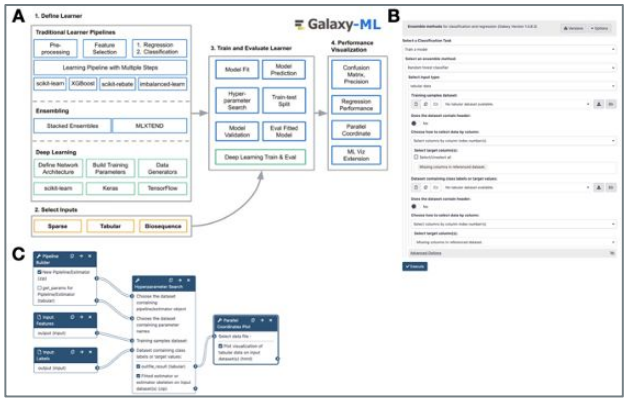
- Plan to sequence and assemble 70,000 high fidelity genomes
- ~1TB input data, 5000 core hours for a typical genome assembly
- Commercial assembly compute cost ~ sequencing cost
- Collaboration: Tools wrapped; Workflows in Galaxy
- Can now use AnVIL Galaxy described above

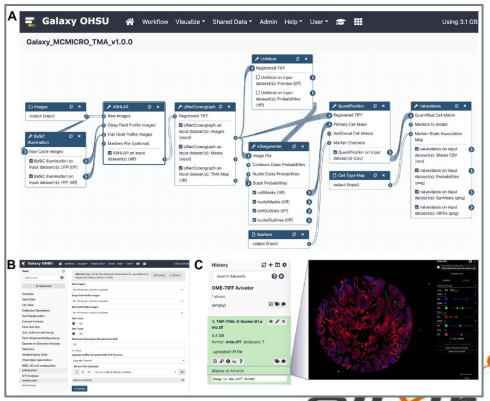


Open Computing: Cancer Informatics

- Distributed analysis of omics and imaging cancer datasets across commercial clouds and institutional HPCs
- Used in U.S. NIH Cancer Moonshot initiative for precision oncology
- Galaxy-ML: a suite of tools for no-code machine and deep learning
- Subcellular microscopy image analysis → 2D and 3D tumor maps

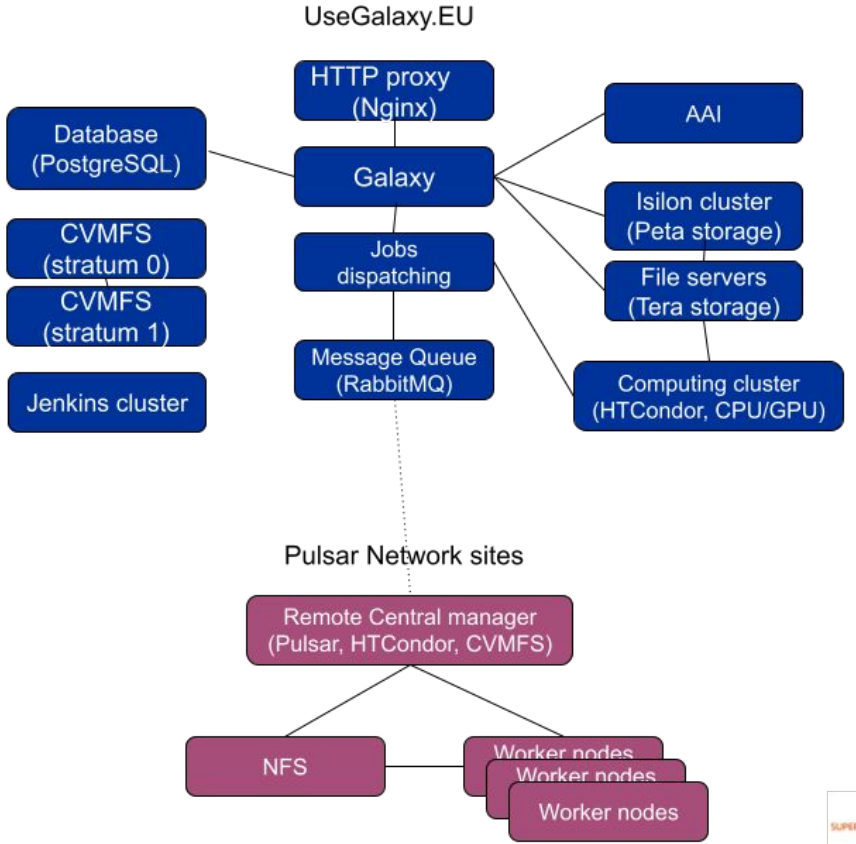


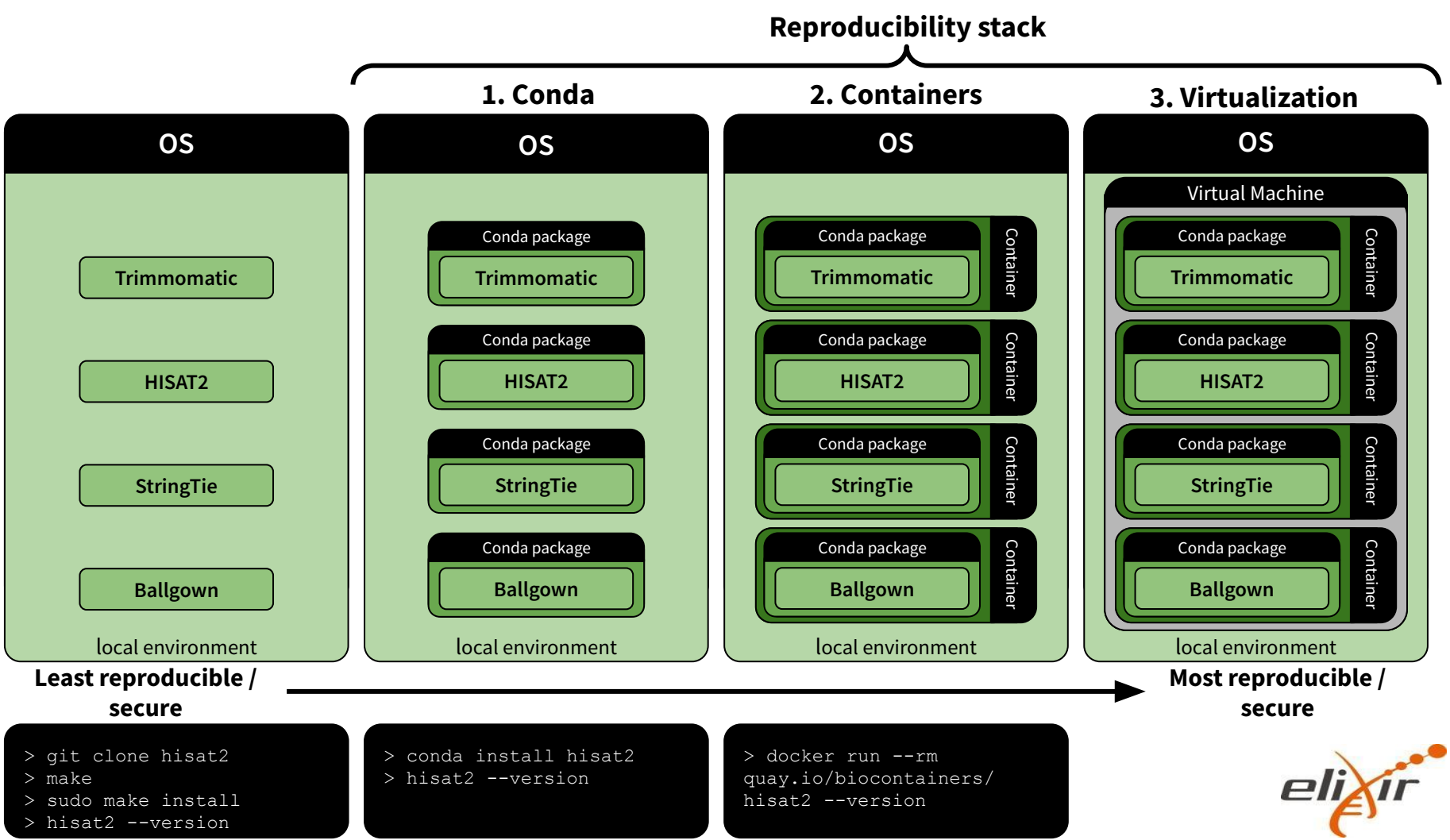




Gu *et al.* (2021) PLoS Computational Biology. https://doi.org/10.1371/journal.pcbi.1009014

Boxes OS.: Centos8
Galaxy: 40 cores, 256GB, 240GB SSD
Database: 40 cores, 256GB, 2TB Raid1 SSD
Jenkins: 40 cores, 128GB, 240GB SSD




Reproducibility stack
1. Conda
2. Containers
3. Virtualization
OS
Trimmomatic
HISAT2
StringTie
Ballgown
local environment
Conda package
Container
Virtual Machine
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