Karolinska Institutet

Pipelines and HPC computing environments

Venkatesh Chellappa, 2022

General Request

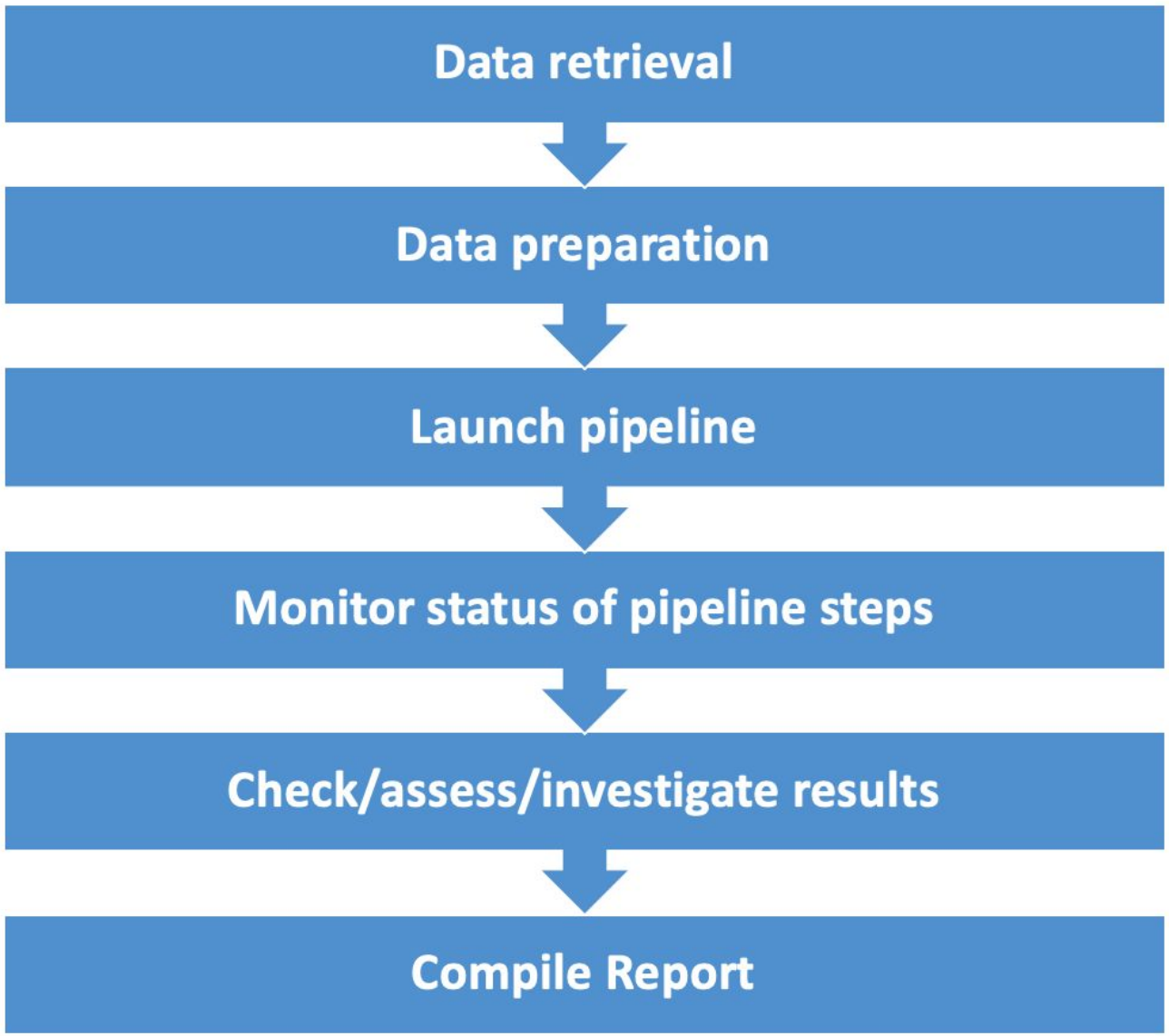
- Some parts of this presentation might contain sensitive copyrighted material.
- Kindly refrain from screenshotting or screen-recording.
- The slides uploaded on the course portal will not have this sensitive information.

This session

- Introduction to NGS data analysis pipelines
- Our Pipeline - LiqBio
- Physical Computer Cluster at MEB, KI
- Cloud Genomics Compute environment @KI - Our Cluster
- SLURM
- Open-source pipelines & platforms for genomic data analysis
- Introduction to the Galaxy Platform

Anatomy of a good NGS data analysis workflow

Karolinska Institutet



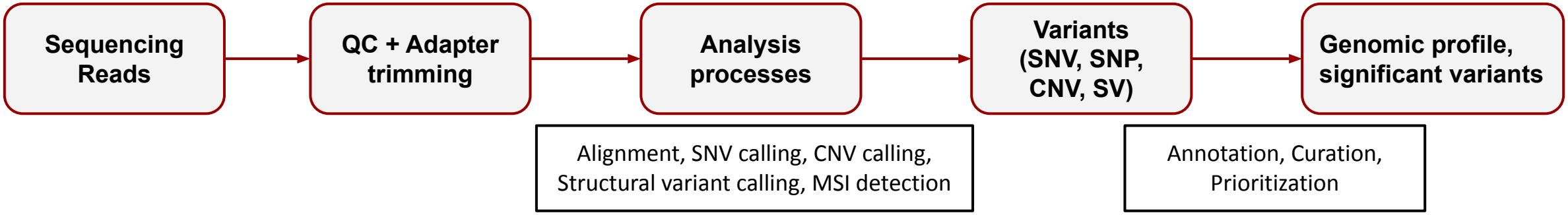
Leaderboard

Andromeda
(autoseq-launcher)

Curator
(autoseq-curator)

NGS data analysis pipeline

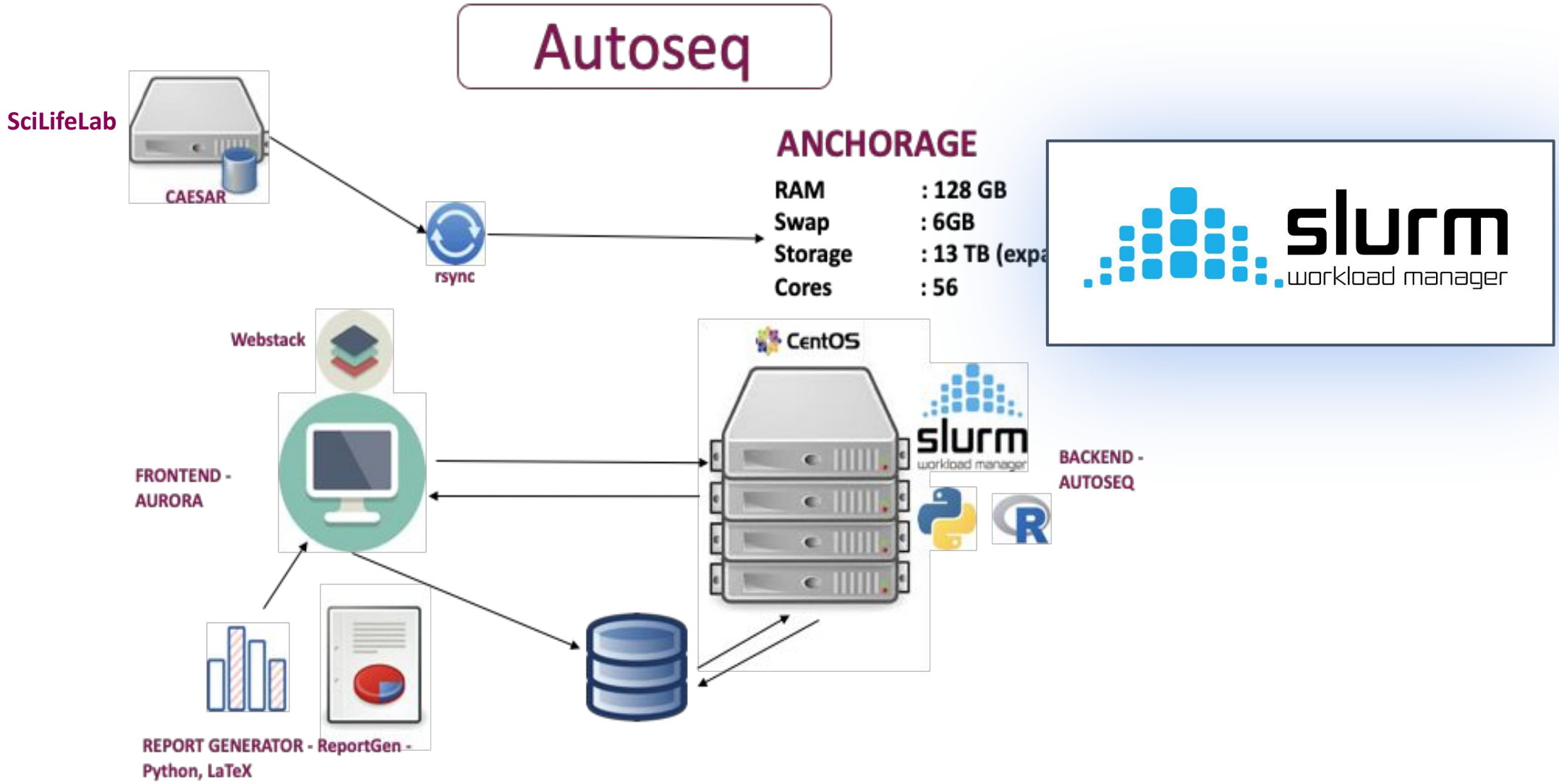
Exome or targeted resequencing data analysis pipeline

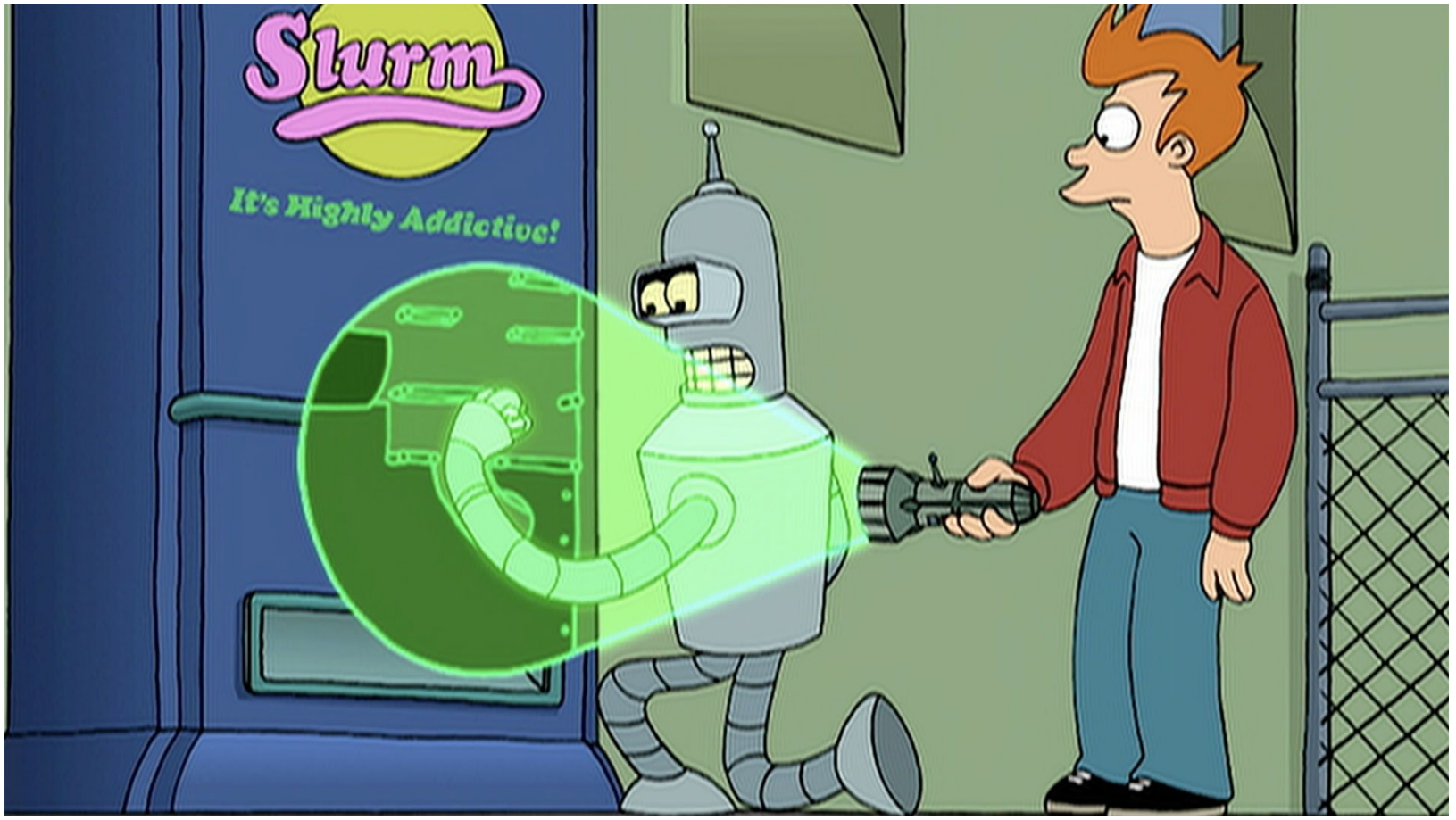


High Performance Computing (HPC) environments

(Physical Cluster at MEB, KI and Cloud Computing cluster at KI)

Physical server setup - overview





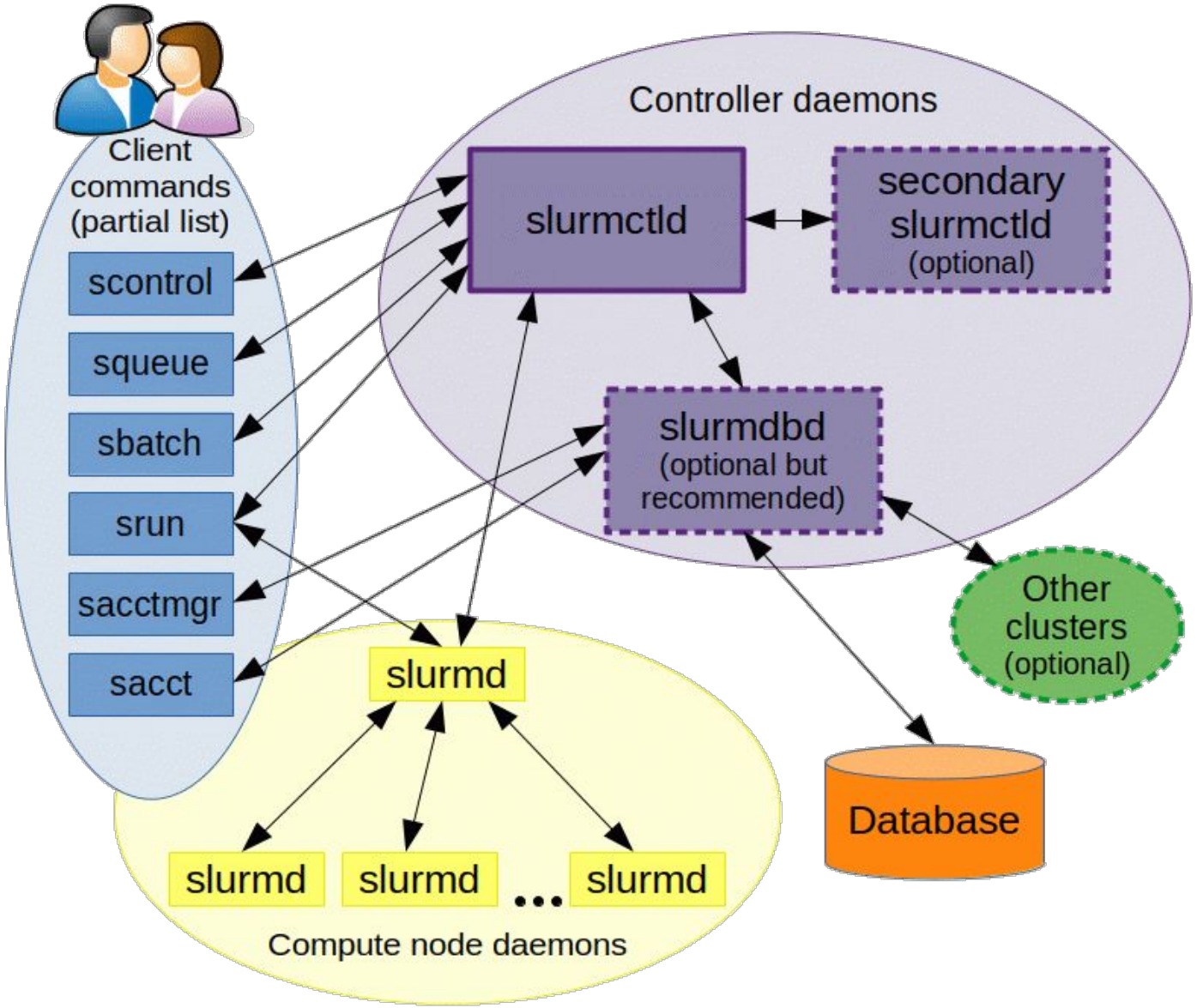
Karolinska Institutet

slurm
workload manager

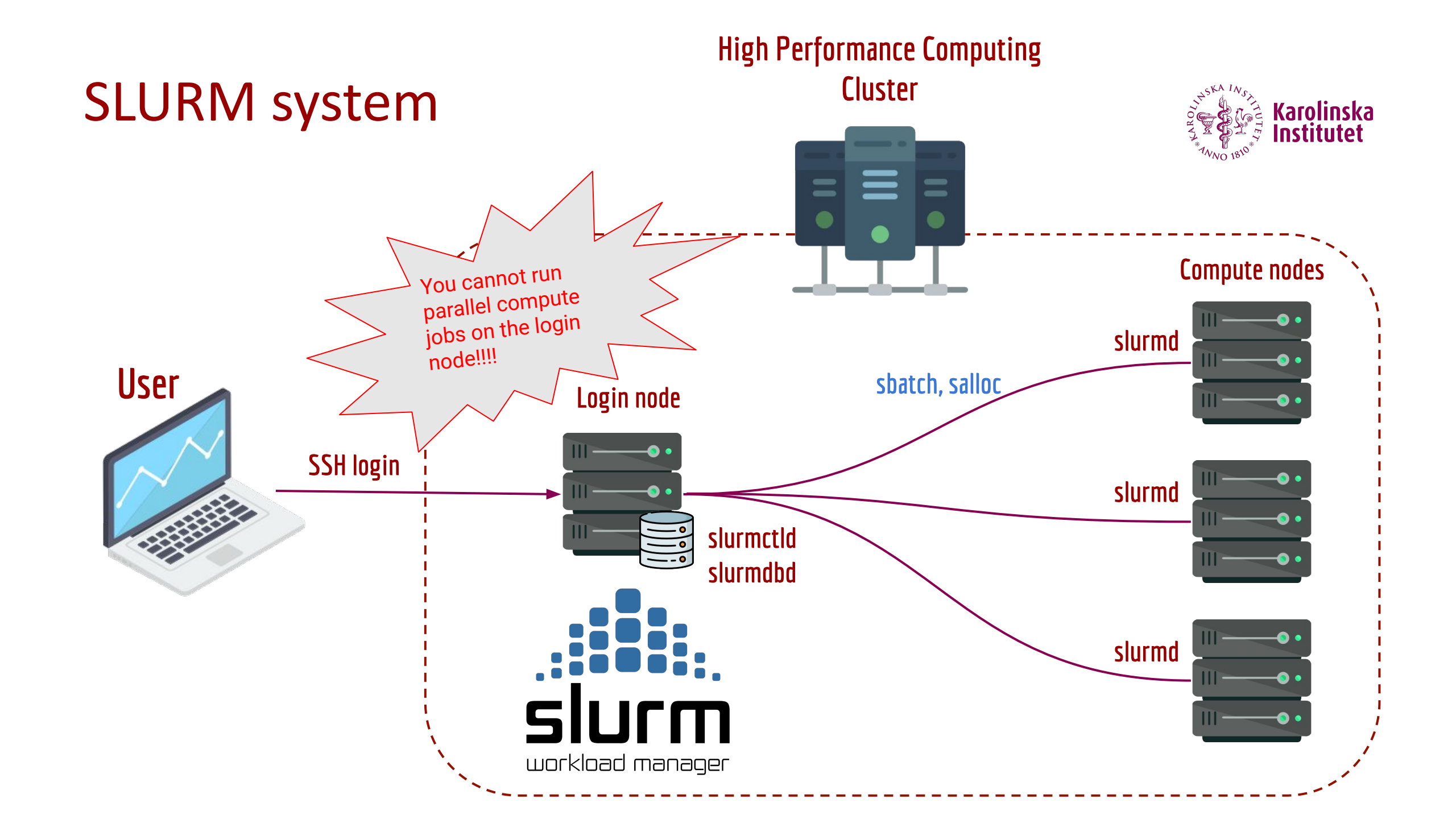
SLURM - Simple Linux Utility for Resource Management

- main utility - workflow manager
- “glue” for all CPUs of a server in parallel computing environment
- makes parallel computing as easy as using a PC
- dynamically manage the usage of resources - CPUs and RAM
- launch multiple jobs in desired sequence or order of priority
- access desired number of CPUs for a dedicated amount of time

SLURM ecosystem



SLURM system



Some open-source pipelines/platforms

Karolinska Institutet

Galaxy PROJECT

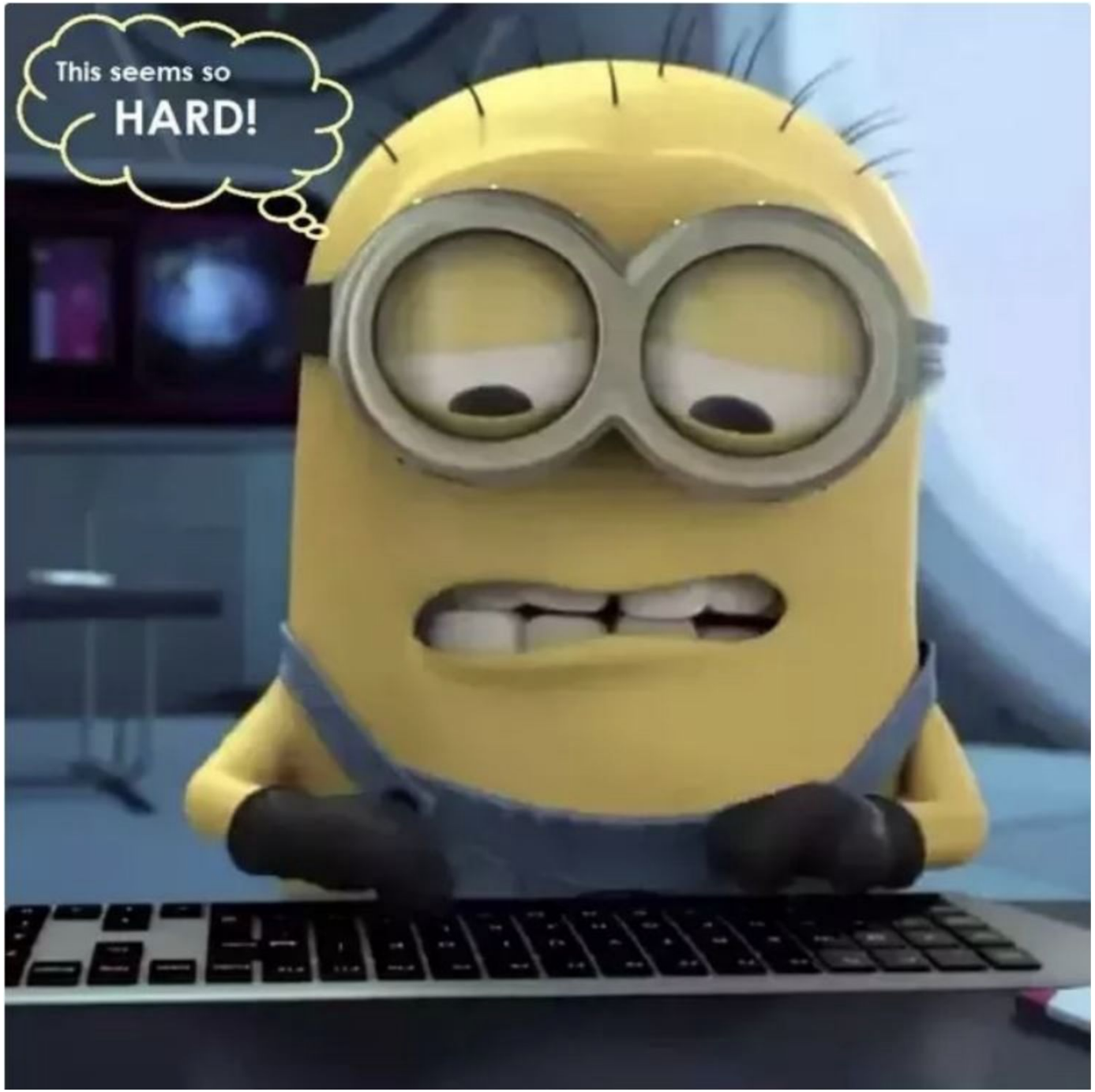
Very thorough and user-friendly but requires intensive training

Open for Innovation KNIME

Closed community, limited usage

nextflow

New platform, many ready-made options but requires programming skills



Karolinska Institutet

Karolinska Institutet

Introduction to Galaxy

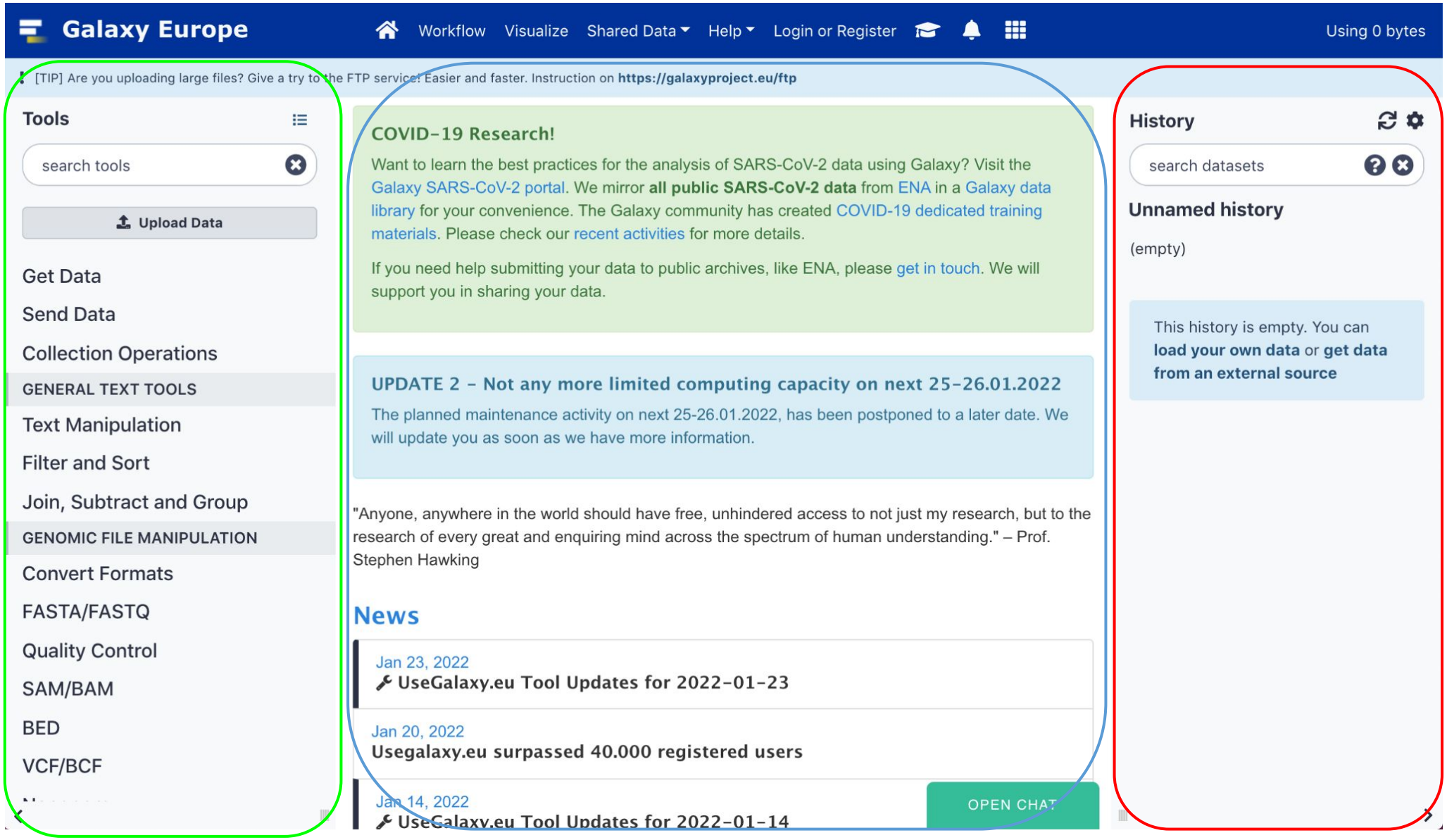
Galaxy

Karolinska Institutet

Data Intensive *analysis* for everyone

- Versatile and reproducible workflows
- **Web** platform
- **Open source** under Academic Free License
- Developed at Penn State, Johns Hopkins, OHSU and Cleveland Clinic with substantial outside contributions

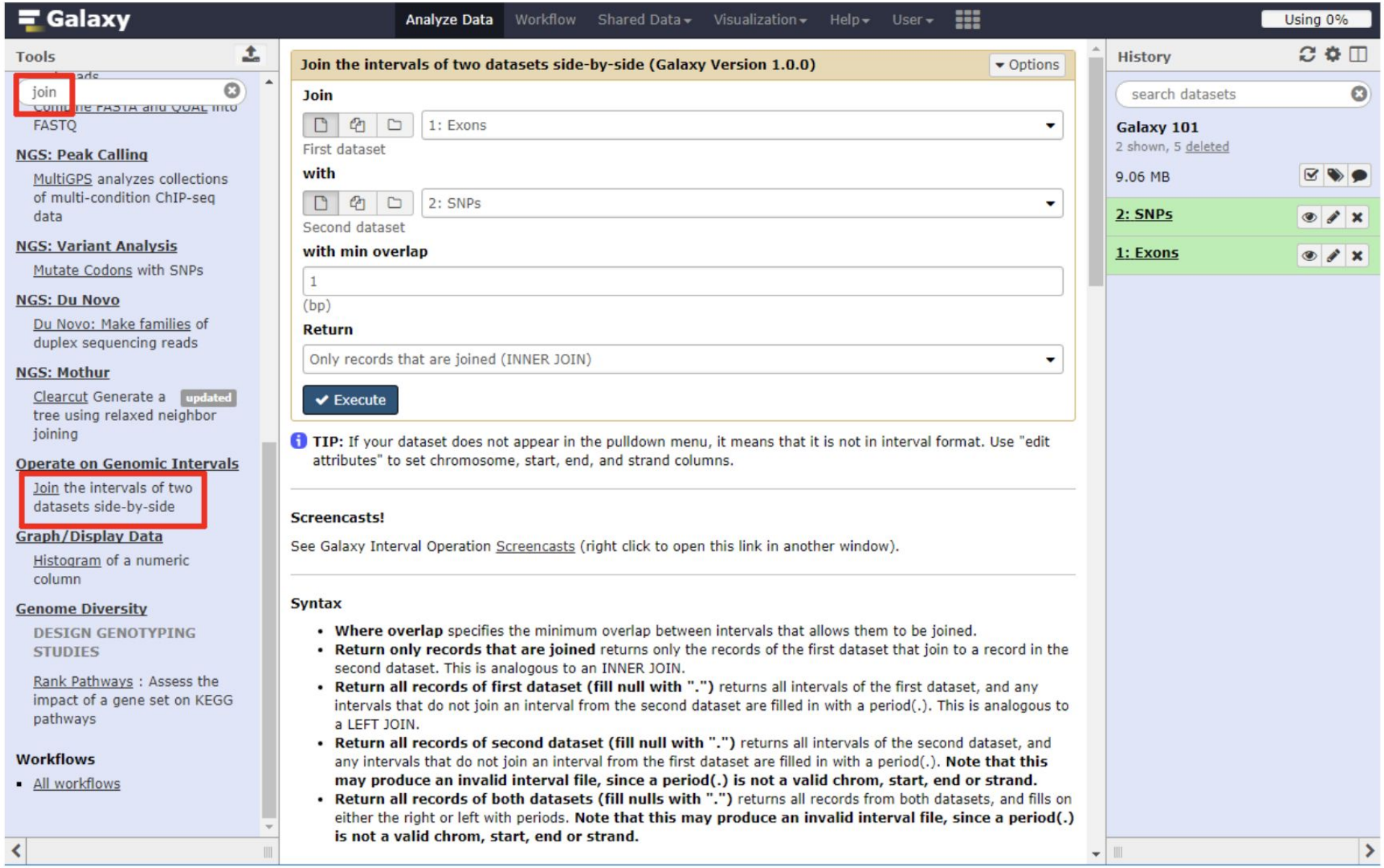
The interface



Top Menu

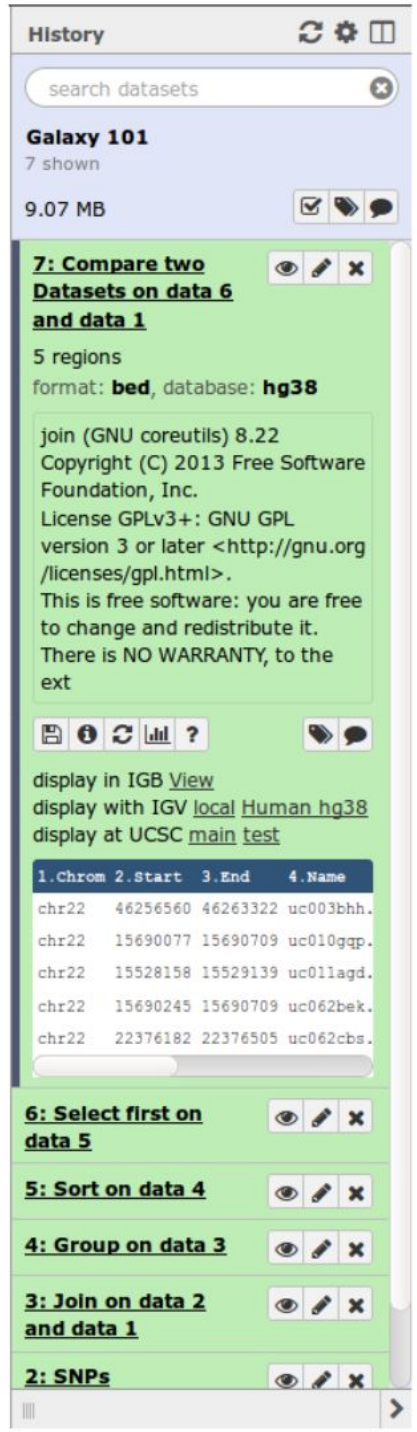
Link	Usage
(or *Analyze Data*)	go back to the homepage
Workflow	access existing workflows or create new one using the editable diagrammatic pipeline
Visualize	create new visualisations and launch Interactive Environments
Shared Data	access data libraries, histories, workflows, visualizations and pages shared with you
Help	links to Galaxy Help Forum (Q&A), Galaxy Community Hub (Wiki), and Interactive Tours
User	your preferences and saved histories, datasets, pages and visualizations

Tools

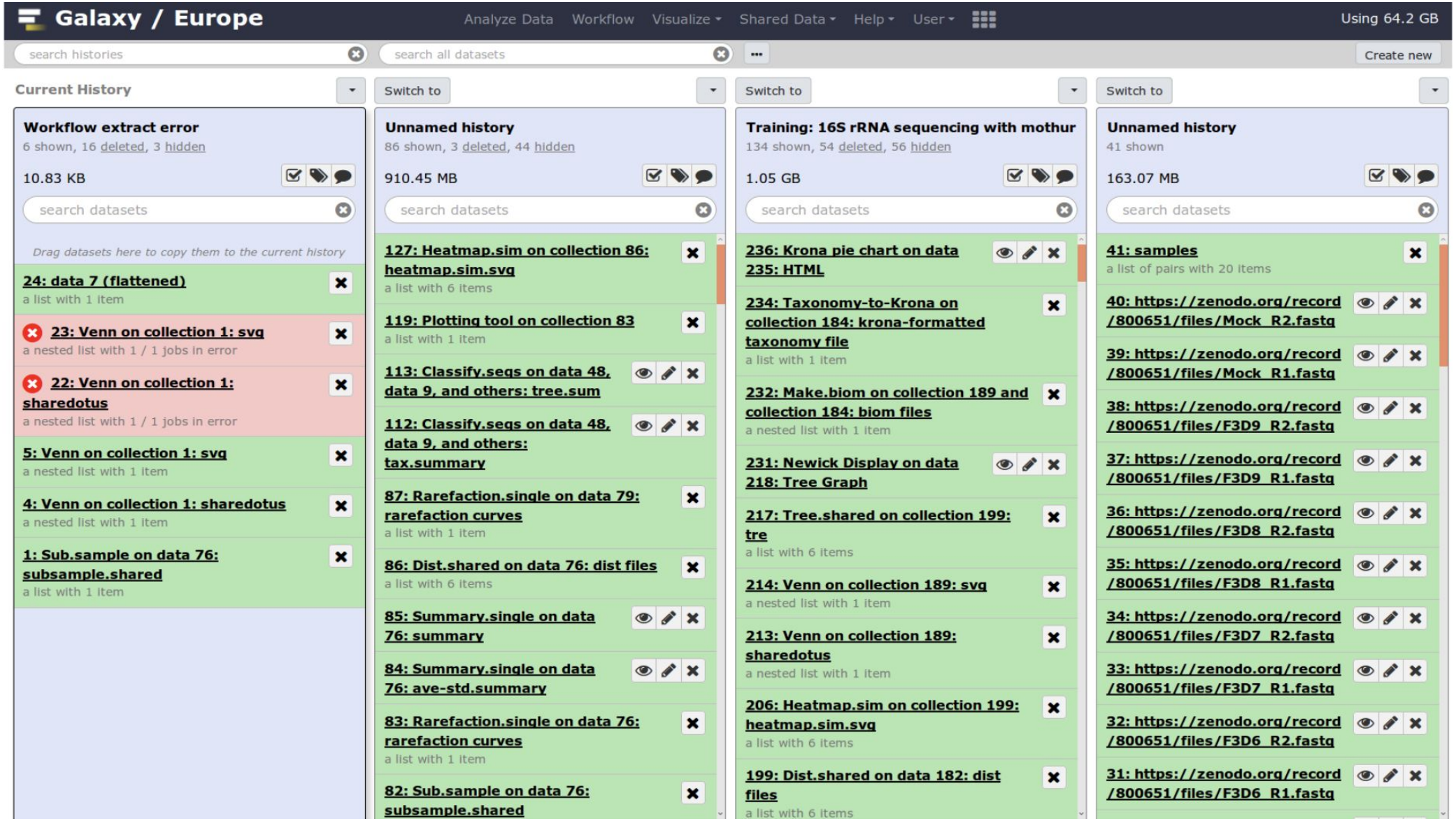

Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 0%
Tools
join
FASTQ
NGS: Peak Calling
MultiGPS analyzes collections of multi-condition ChIP-seq data
NGS: Variant Analysis
Mutate Codons with SNPs
NGS: Du Novo
Du Novo: Make families of duplex sequencing reads
NGS: Mothur
Clearcut Generate a tree using relaxed neighbor joining
updated
Operate on Genomic Intervals
Join the intervals of two datasets side-by-side
Graph/Display Data
Histogram of a numeric column
Genome Diversity
DESIGN GENOTYPING STUDIES
Rank Pathways : Assess the impact of a gene set on KEGG pathways
Workflows
All workflows
Join the intervals of two datasets side-by-side (Galaxy Version 1.0.0)
Options
Join
1: Exons
First dataset
with
2: SNPs
Second dataset
with min overlap
1
(bp)
Return
Only records that are joined (INNER JOIN)
Execute
TIP: If your dataset does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.
Screencasts!
See Galaxy Interval Operation Screencasts (right click to open this link in another window).
Syntax
Where overlap specifies the minimum overlap between intervals that allows them to be joined.
Return only records that are joined returns only the records of the first dataset that join to a record in the second dataset. This is analogous to an INNER JOIN.
Return all records of first dataset (fill null with ".") returns all intervals of the first dataset, and any intervals that do not join an interval from the second dataset are filled in with a period(.). This is analogous to a LEFT JOIN.
Return all records of second dataset (fill null with ".") returns all intervals of the second dataset, and any intervals that do not join an interval from the first dataset are filled in with a period(.). Note that this may produce an invalid interval file, since a period(.) is not a valid chrom, start, end or strand.
Return all records of both datasets (fill nulls with ".") returns all records from both datasets, and fills on either the right or left with periods. Note that this may produce an invalid interval file, since a period(.) is not a valid chrom, start, end or strand.
History
search datasets
Galaxy 101
2 shown, 5 deleted
9.06 MB
2: SNPs
1: Exons

History

- Location of all analyses
  - collects all datasets produced by tools
  - collects all operations performed on the data
- For each dataset (the heart of Galaxy's reproducibility), the history tracks
  - name, format, size, creation time, datatype-specific metadata
  - tool id, version, inputs, parameters
  - standard output (`stdout`) and error (`stderr`)
  - state (waiting, running, success, failed)
  - hidden, deleted, purged

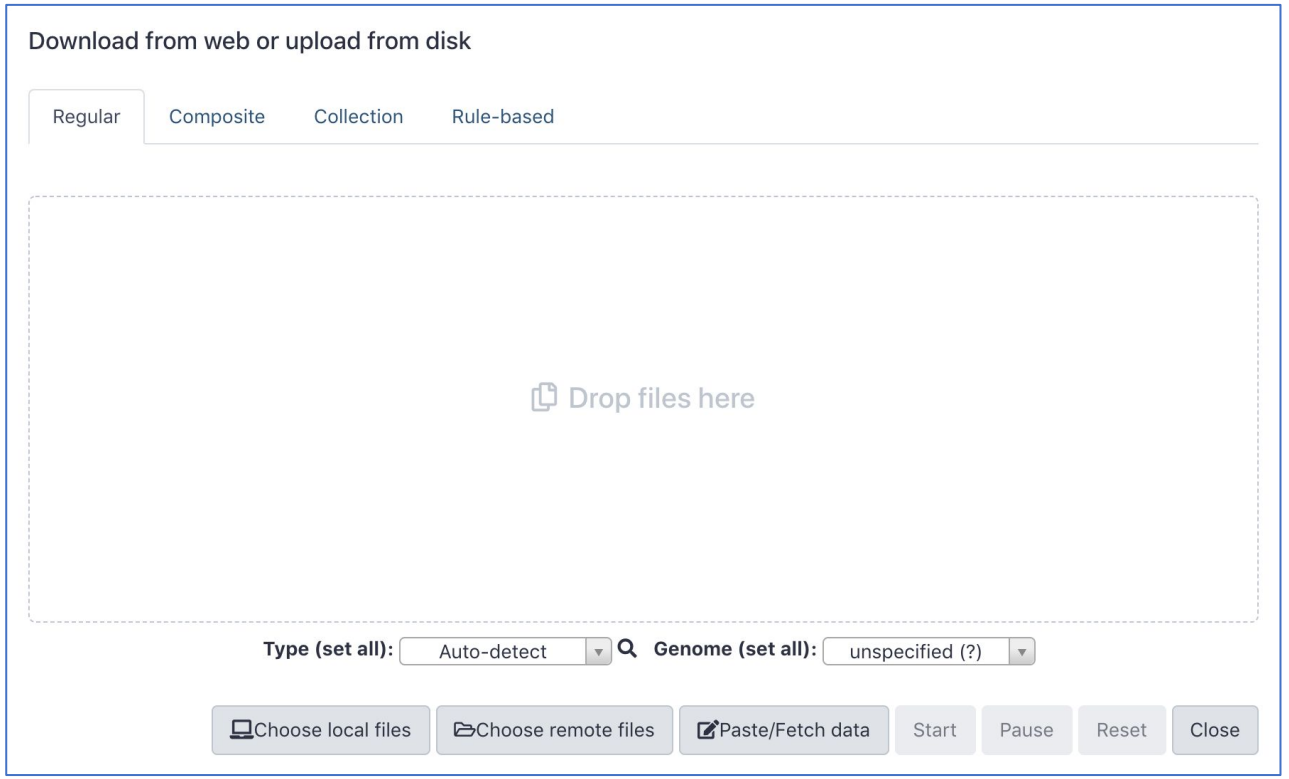


Multiple histories



Importing data

- Copy/paste some text
- Upload files from your local computer
- Upload data from an internet URL
- Upload data from online databases: UCSC, BioMart, ENCODE, modENCODE, Flymine etc.
- Import from Shared Data (libraries, histories, pages)
- Upload data from FTP

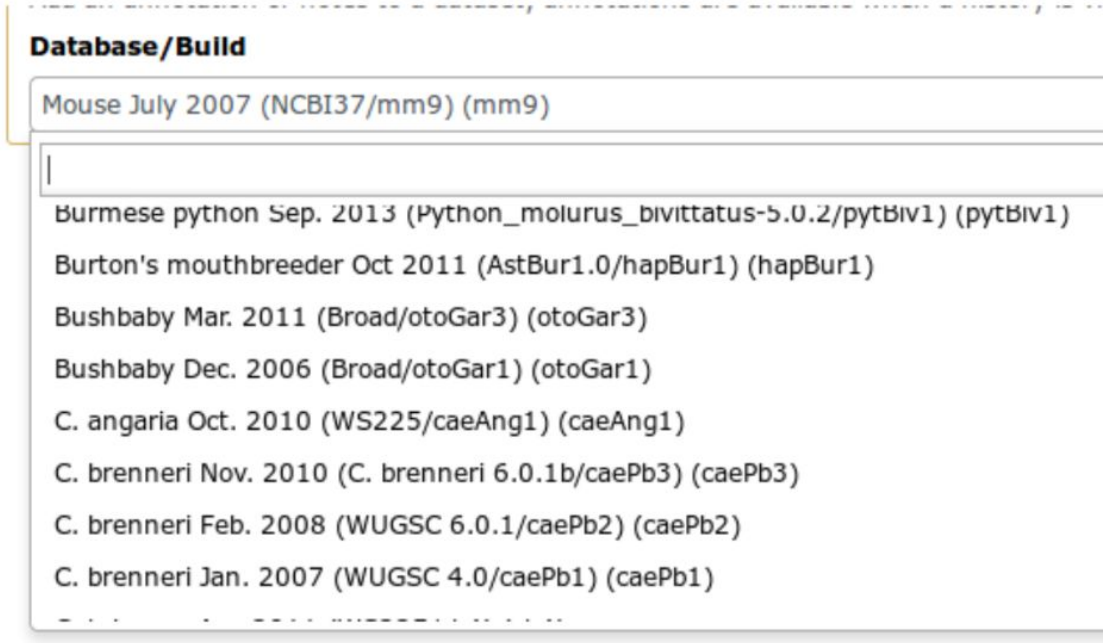


Loading reference genomes

Karolinska Institutet

Example: reference Genome

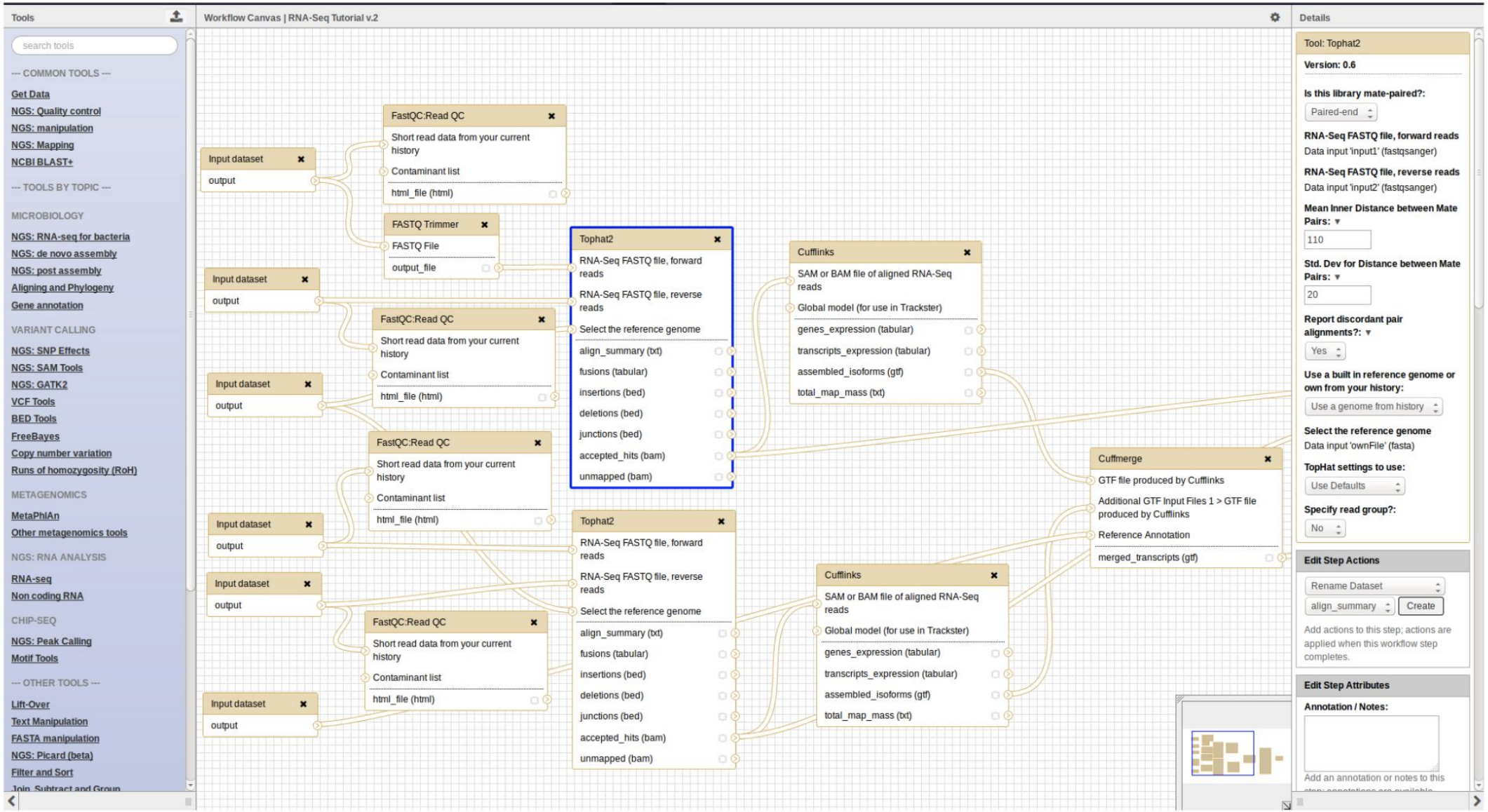
- Genome build specifies which genome assembly a dataset is associated with
  - e.g. mm10, hg38…
- Can be assigned by a tool or by the user
- Users can create custom genome builds
- New builds can be added by the admin



Workflow

Karolinska Institutet

Workflow Editor



Advantages of creating and saving workflows

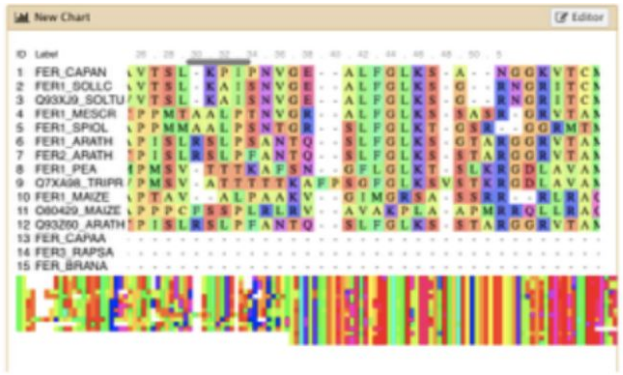
Karolinska Institutet

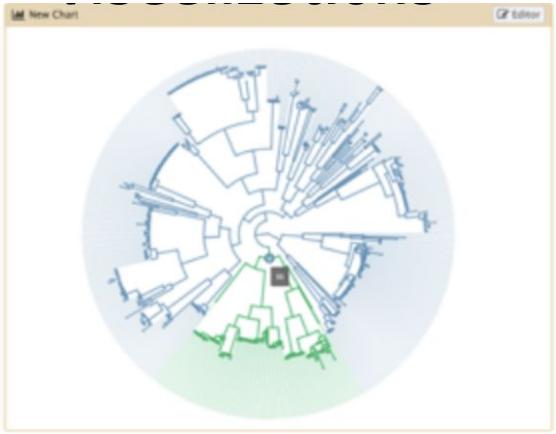
- **Re-run** the same analysis on different input data sets
- **Change parameters** before re-running a similar analysis
- Make use of the workflow job **scheduling**
  - jobs are submitted as soon as their inputs are ready
- Create **sub**-workflows: a workflow inside another workflow
- **Share** workflows for publication and with the community

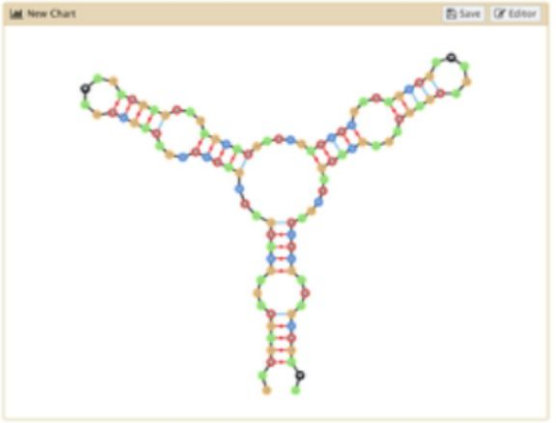
Visualization in Galaxy

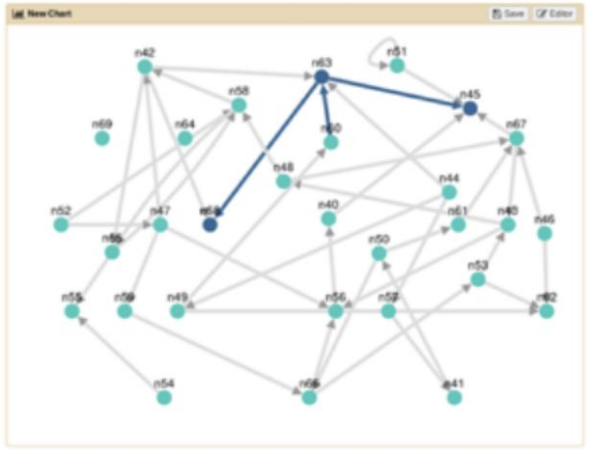
Karolinska Institutet

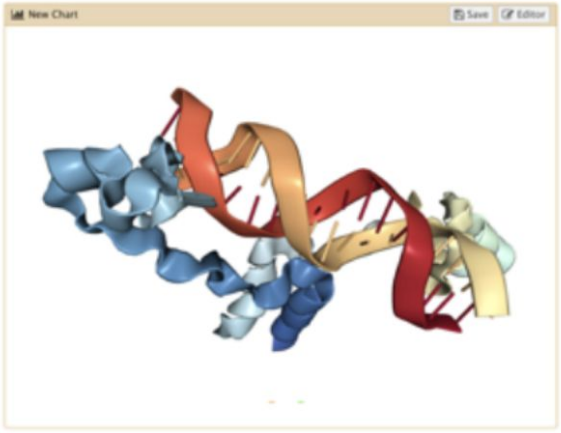
Charts in Galaxy

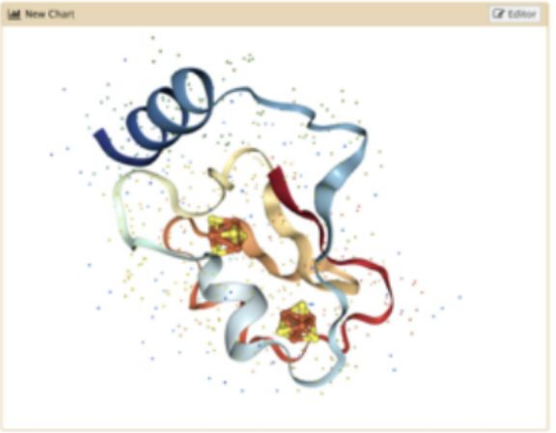




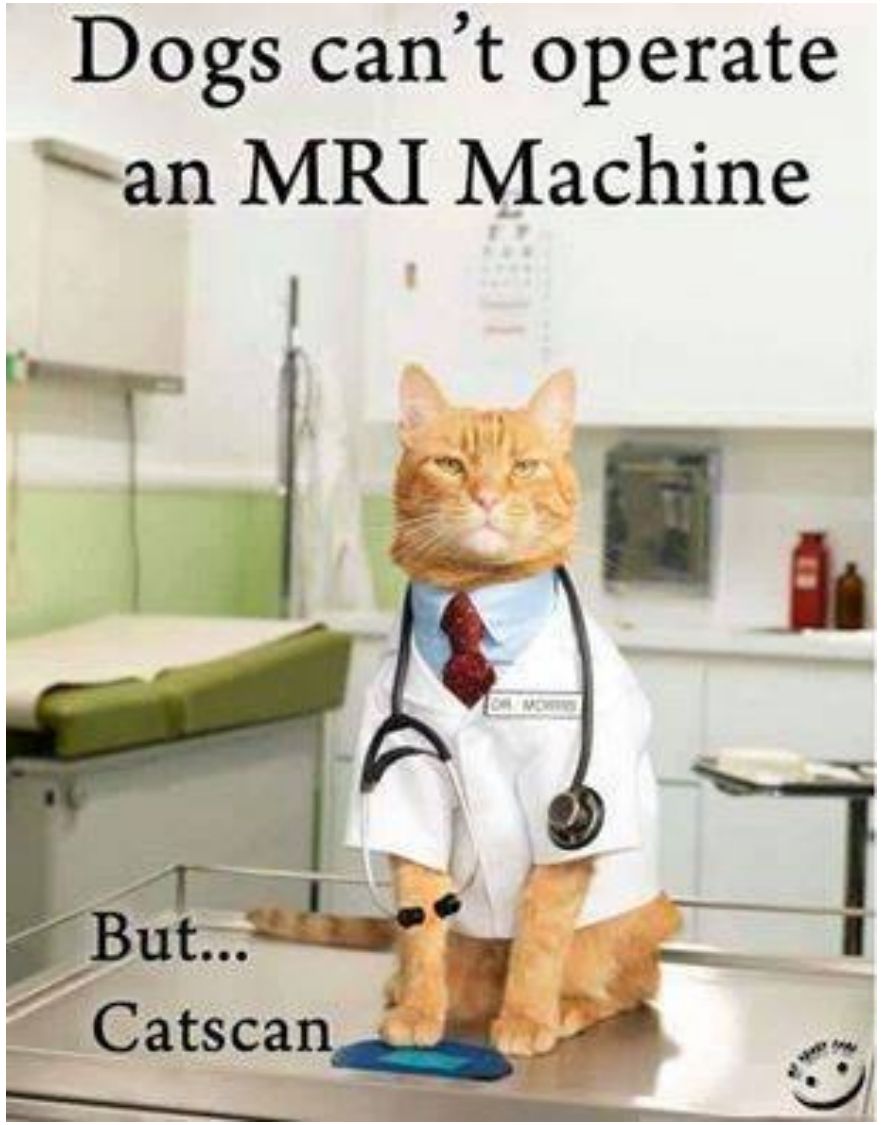








Questions



Credits

Content for several slides on Galaxy Platform were borrowed from the Galaxy Tutorials.

Galaxy Training Network (https://training.galaxyproject.org/)