



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

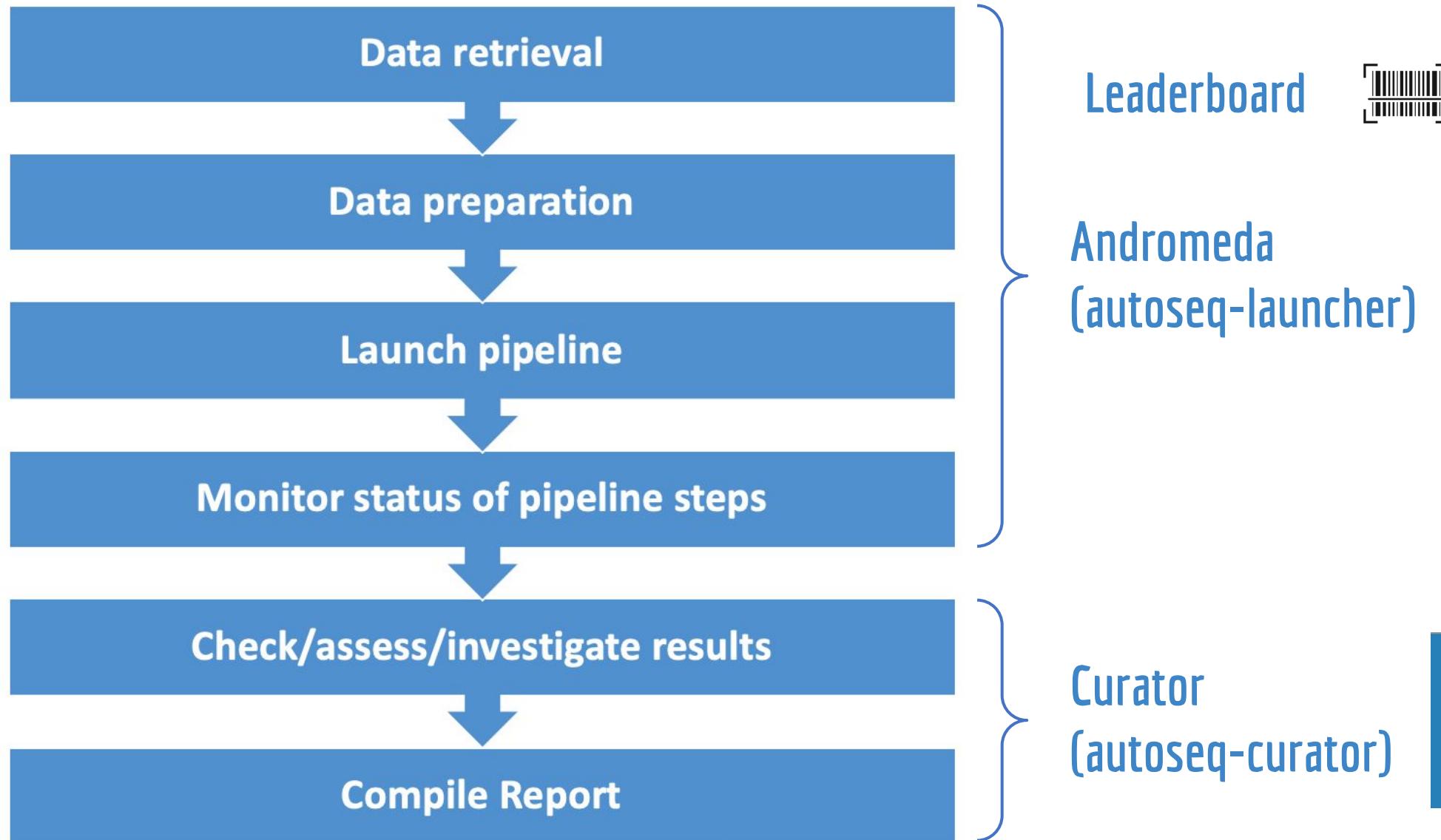
- Introduction to NGS data analysis pipelines
- Our Pipeline - LiqBio
- Physical Computer Cluster at MEB, KI
- Cloud Genomics Compute environment @KI - Our Cluster
- SLURM
- Open-source pipelines & platforms for genomic data analysis
- Introduction to the Galaxy Platform



Anatomy of a good NGS data analysis workflow



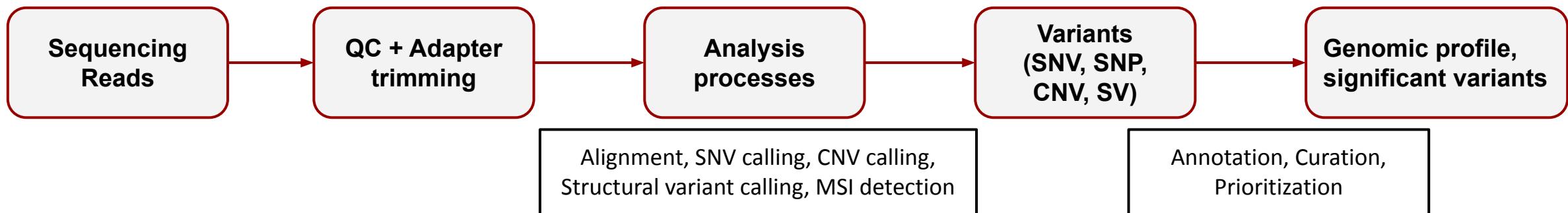
Karolinska
Institutet



NGS data analysis pipeline



Exome or targeted resequencing data analysis pipeline



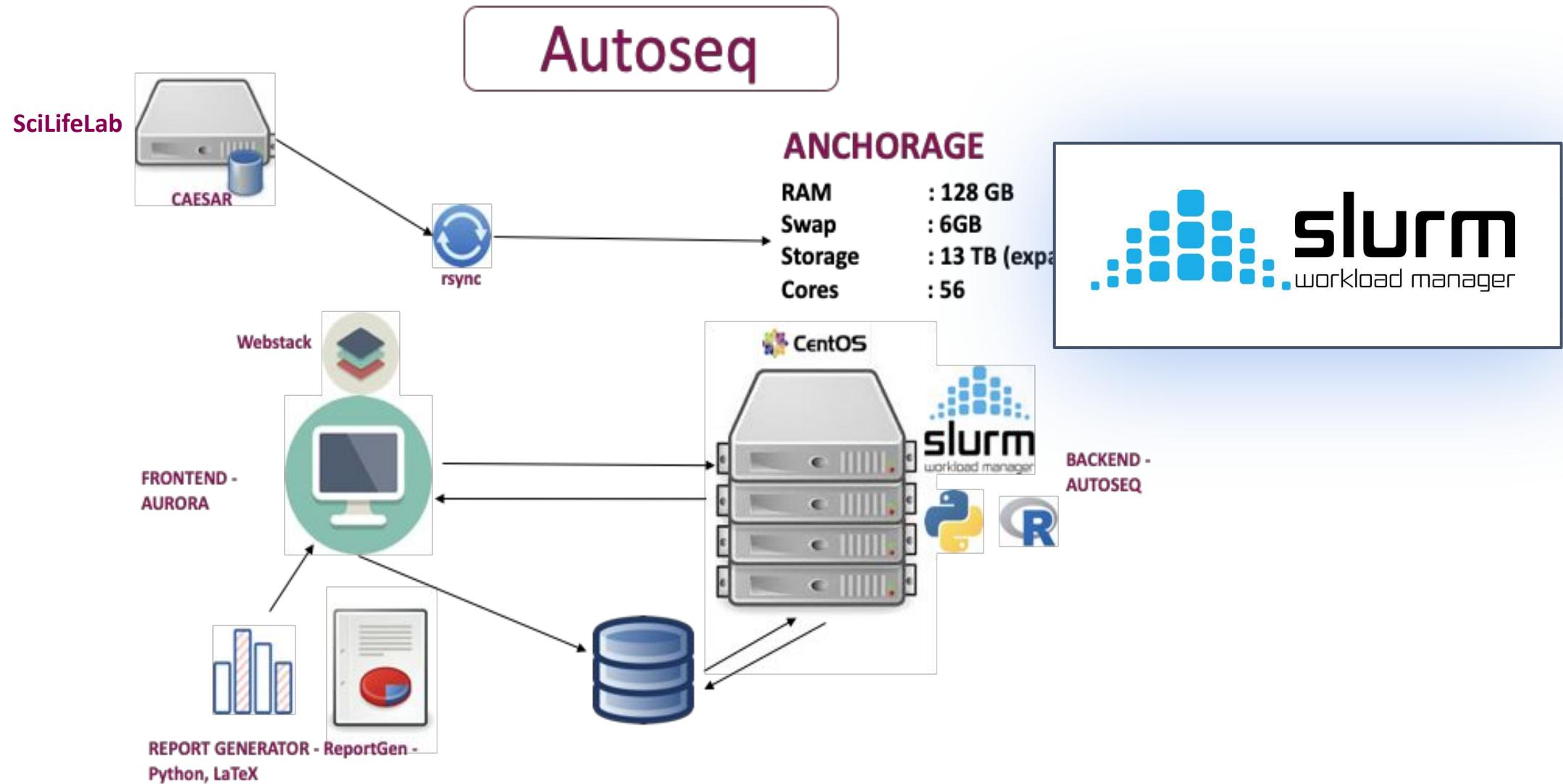
High Performance Computing (HPC) environments

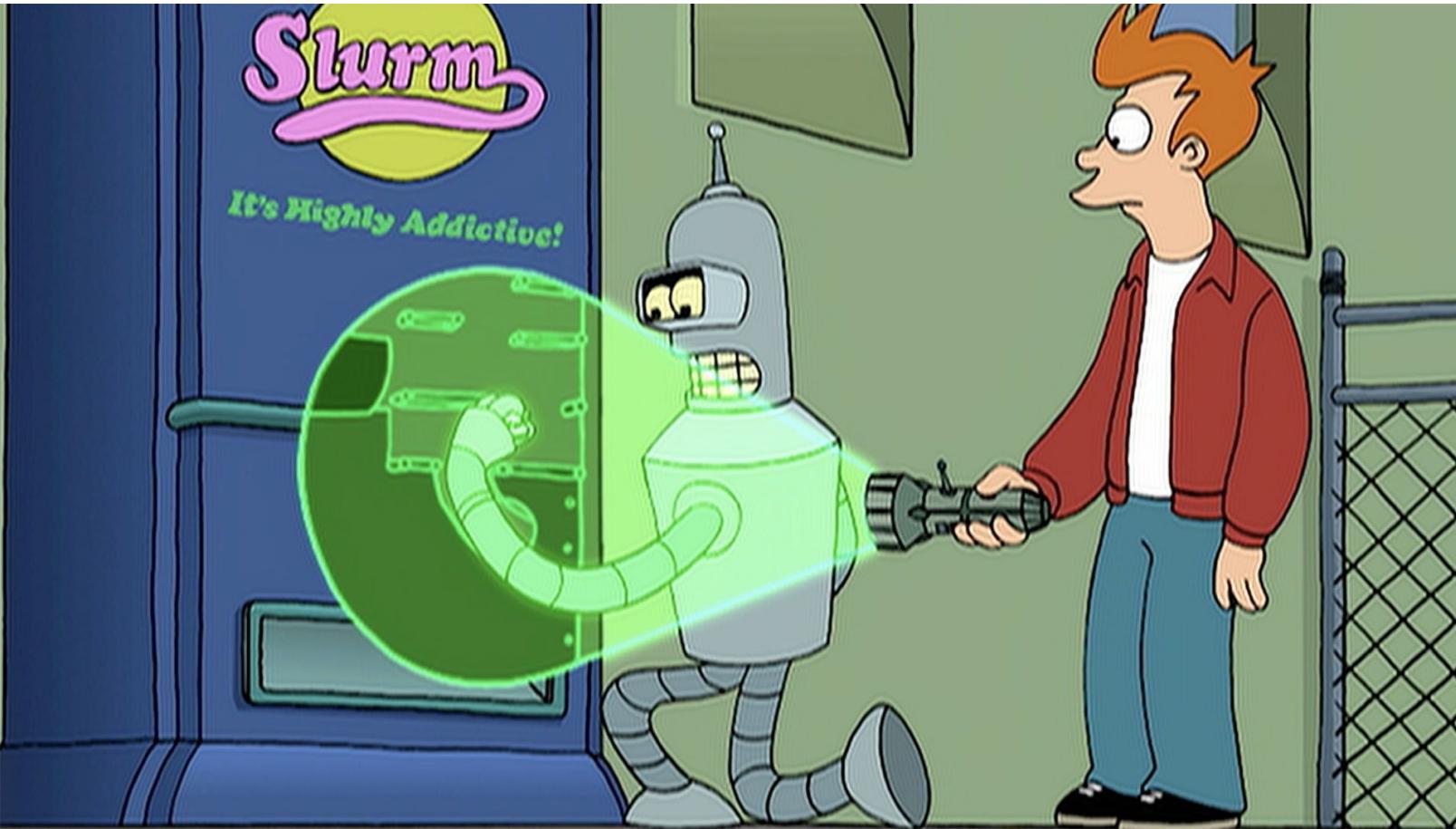
(Physical Cluster at MEB, KI and Cloud Computing cluster at KI)

Physical server setup - overview



Karolinska
Institutet





Karolinska
Institutet



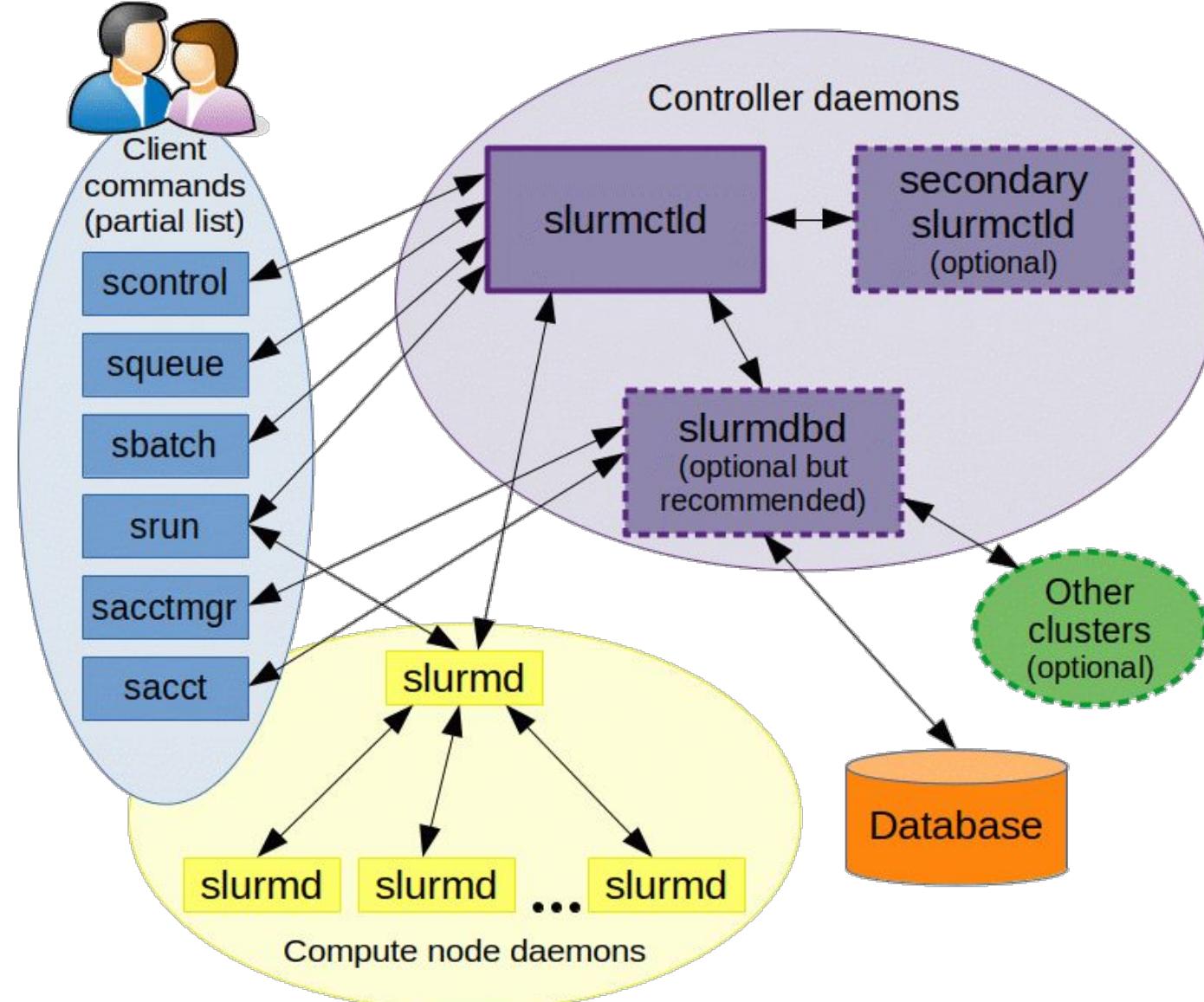
SLURM - Simple Linux Utility for Resource Management



- 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

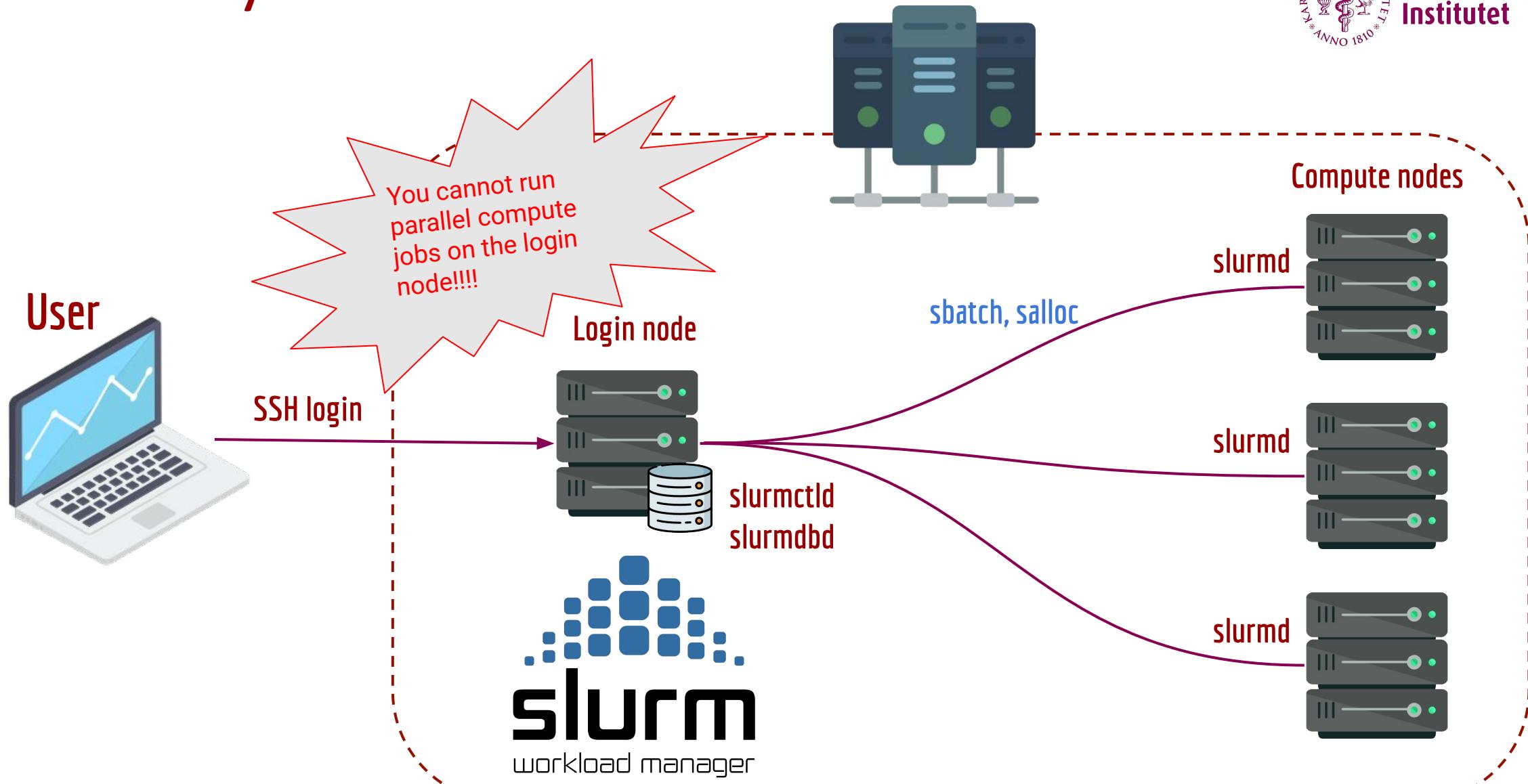


SLURM ecosystem



High Performance Computing Cluster

SLURM system



Some open-source pipelines/platforms



Very thorough and user-friendly but requires intensive training



Closed community, limited usage



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



The screenshot shows the Galaxy Europe web interface. The top navigation bar includes links for Workflow, Visualize, Shared Data, Help, Login or Register, and a user icon. A tip message at the top left suggests using the FTP service for large files. The main content area features a green-highlighted sidebar with various tools like search tools, upload data, get data, send data, collection operations, general text tools (selected), text manipulation, filter and sort, join, subtract and group, genomic file manipulation (selected), convert formats, FASTA/FASTQ, quality control, SAM/BAM, BED, and VCF/BCF. A blue-highlighted section titled "COVID-19 Research!" provides information about SARS-CoV-2 analysis and training materials. A blue-highlighted box contains an "UPDATE 2" notice about postponed maintenance. A red-highlighted sidebar on the right shows an empty history section with options to load data or get data from external sources.

[TIP] Are you uploading large files? Give a try to the FTP service! Easier and faster. Instruction on <https://galaxyproject.eu/ftp>

Tools

- search tools
- Upload Data

Get Data

Send Data

Collection Operations

GENERAL TEXT TOOLS (selected)

Text Manipulation

Filter and Sort

Join, Subtract and Group

GENOMIC FILE MANIPULATION (selected)

Convert Formats

FASTA/FASTQ

Quality Control

SAM/BAM

BED

VCF/BCF

COVID-19 Research!

Want to learn the best practices for the analysis of SARS-CoV-2 data using Galaxy? Visit the [Galaxy SARS-CoV-2 portal](#). We mirror **all public SARS-CoV-2 data** from ENA in a [Galaxy data library](#) for your convenience. The Galaxy community has created [COVID-19 dedicated training materials](#). Please check our [recent activities](#) for more details.

If you need help submitting your data to public archives, like ENA, please [get in touch](#). We will support you in sharing your data.

UPDATE 2 – Not any more limited computing capacity on next 25–26.01.2022

The planned maintenance activity on next 25–26.01.2022, has been postponed to a later date. We will update you as soon as we have more information.

"Anyone, anywhere in the world should have free, unhindered access to not just my research, but to the research of every great and enquiring mind across the spectrum of human understanding." – Prof. Stephen Hawking

News

Jan 23, 2022 [UseGalaxy.eu Tool Updates for 2022-01-23](#)

Jan 20, 2022 [Usegalaxy.eu surpassed 40.000 registered users](#)

Jan 14, 2022 [UseGalaxy.eu Tool Updates for 2022-01-14](#)

OPEN CHAT

Using 0 bytes

History

- search datasets

Unnamed history

(empty)

This history is empty. You can [load your own data](#) or [get data from an external source](#)

Top Menu



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

Tools

Galaxy Analyze Data Workflow Shared Data Visualization Help User Using 0%

Tools

join (highlighted with a red box)

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

Operate on Genomic Intervals

Join the intervals of two datasets side-by-side (highlighted with a red box)

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)

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

A screenshot of the Galaxy History interface. At the top, there is a search bar labeled "search datasets" and a section titled "Galaxy 101" showing "7 shown" datasets of size "9.07 MB". Below this, a list of analysis steps is displayed:

- 7: Compare two Datasets on data 6 and data 1
5 regions
format: bed, database: hg38
join (GNU coreutils) 8.22
Copyright (C) 2013 Free Software Foundation, Inc.
License GPLv3+: GNU GPL version 3 or later <<http://gnu.org/licenses/gpl.html>>. This is free software: you are free to change and redistribute it. There is NO WARRANTY, to the ext
- display in IGB View
display with IGV local Human hg38
display at UCSC main test
- 1.Chrom 2.Start 3.End 4.Name
chr22 46256560 46263322 uc003bhh.
chr22 15690077 15690709 uc010gqp.
chr22 15528158 15529139 uc011agd.
chr22 15690245 15690709 uc062bek.
chr22 22376182 22376505 uc062cbs.
- 6: Select first on data 5
- 5: Sort on data 4
- 4: Group on data 3
- 3: Join on data 2 and data 1
- 2: SNPs

Multiple histories

Galaxy / Europe Analyze Data Workflow Visualize Shared Data Help User Using 64.2 GB

search histories search all datasets Create new

Current History Switch to Switch to Switch to Switch to

Workflow extract error
6 shown, 16 deleted, 3 hidden
10.83 KB

search datasets

Drag datasets here to copy them to the current history

24: data 7 (flattened)
a list with 1 item

23: Venn on collection 1: svg
a nested list with 1 / 1 jobs in error

22: Venn on collection 1: sharedotus
a nested list with 1 / 1 jobs in error

5: Venn on collection 1: svg
a nested list with 1 item

4: Venn on collection 1: sharedotus
a nested list with 1 item

1: Sub.sample on data 76: subsample.shared
a list with 1 item

Unnamed history
86 shown, 3 deleted, 44 hidden
910.45 MB

search datasets

127: Heatmap.sim on collection 86: heatmap.sim.svg
a list with 6 items

119: Plotting tool on collection 83
a list with 1 item

113: Classify.seqs on data 48, data 9, and others: tree.sum

112: Classify.seqs on data 48, data 9, and others: tax.summary

87: Rarefaction.single on data 79: rarefaction.curves
a list with 1 item

86: Dist.shared on data 76: dist files
a list with 6 items

85: Summary.single on data 76: summary

84: Summary.single on data 76: ave-std.summary

83: Rarefaction.single on data 76: rarefaction.curves
a list with 1 item

82: Sub.sample on data 76: subsample.shared

Training: 16S rRNA sequencing with mothur
134 shown, 54 deleted, 56 hidden
1.05 GB

search datasets

236: Krona pie chart on data

235: HTML

234: Taxonomy-to-Krona on collection 184: krona-formatted taxonomy file
a list with 1 item

232: Make.biom on collection 189 and collection 184: biom files
a nested list with 1 item

231: Newick Display on data

218: Tree Graph

217: Tree.shared on collection 199: tre
a list with 6 items

214: Venn on collection 189: svg
a nested list with 1 item

213: Venn on collection 189: sharedotus
a nested list with 1 item

206: Heatmap.sim on collection 199: heatmap.sim.svg
a list with 6 items

199: Dist.shared on data 182: dist files
a list with 6 items

Unnamed history
41 shown
163.07 MB

search datasets

41: samples
a list of pairs with 20 items

40: https://zenodo.org/record/800651/files/Mock_R2.fastq

39: https://zenodo.org/record/800651/files/Mock_R1.fastq

38: https://zenodo.org/record/800651/files/F3D9_R2.fastq

37: https://zenodo.org/record/800651/files/F3D9_R1.fastq

36: https://zenodo.org/record/800651/files/F3D8_R2.fastq

35: https://zenodo.org/record/800651/files/F3D8_R1.fastq

34: https://zenodo.org/record/800651/files/F3D7_R2.fastq

33: https://zenodo.org/record/800651/files/F3D7_R1.fastq

32: https://zenodo.org/record/800651/files/F3D6_R2.fastq

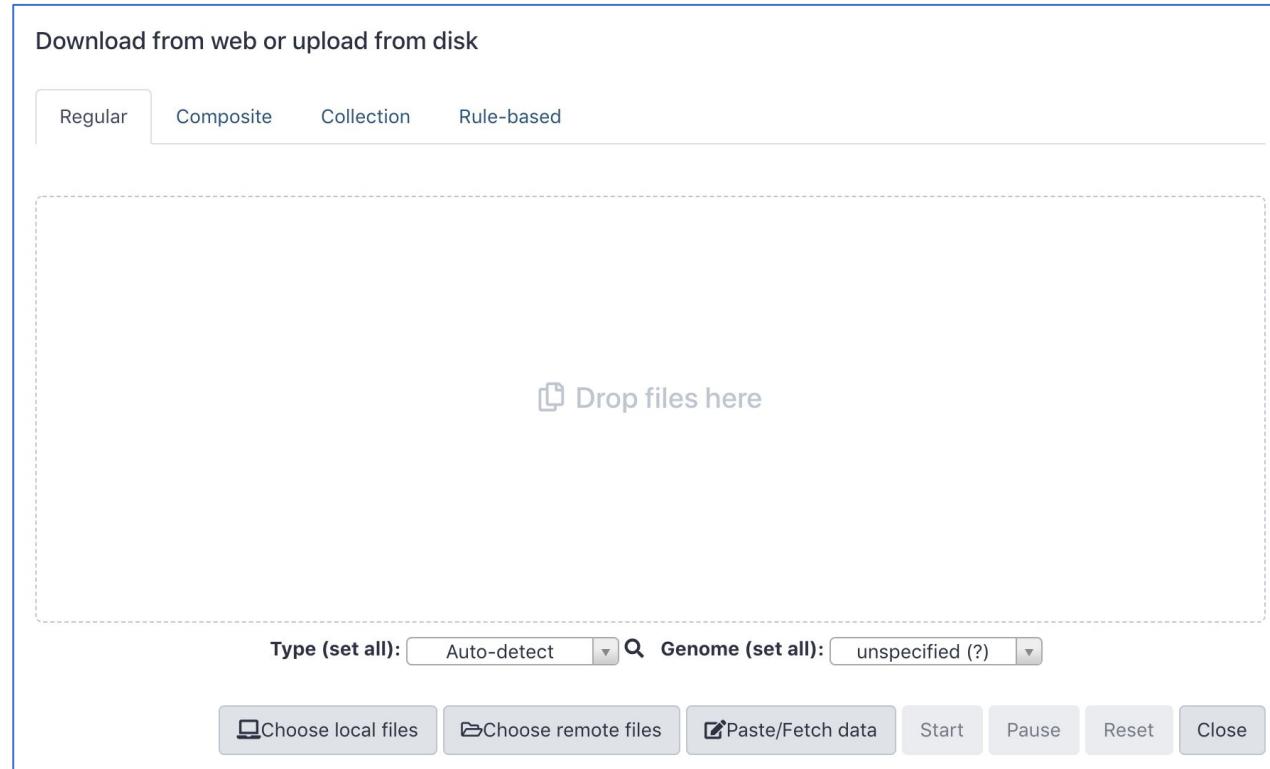
31: https://zenodo.org/record/800651/files/F3D6_R1.fastq

Importing data



Karolinska
Institutet

- 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

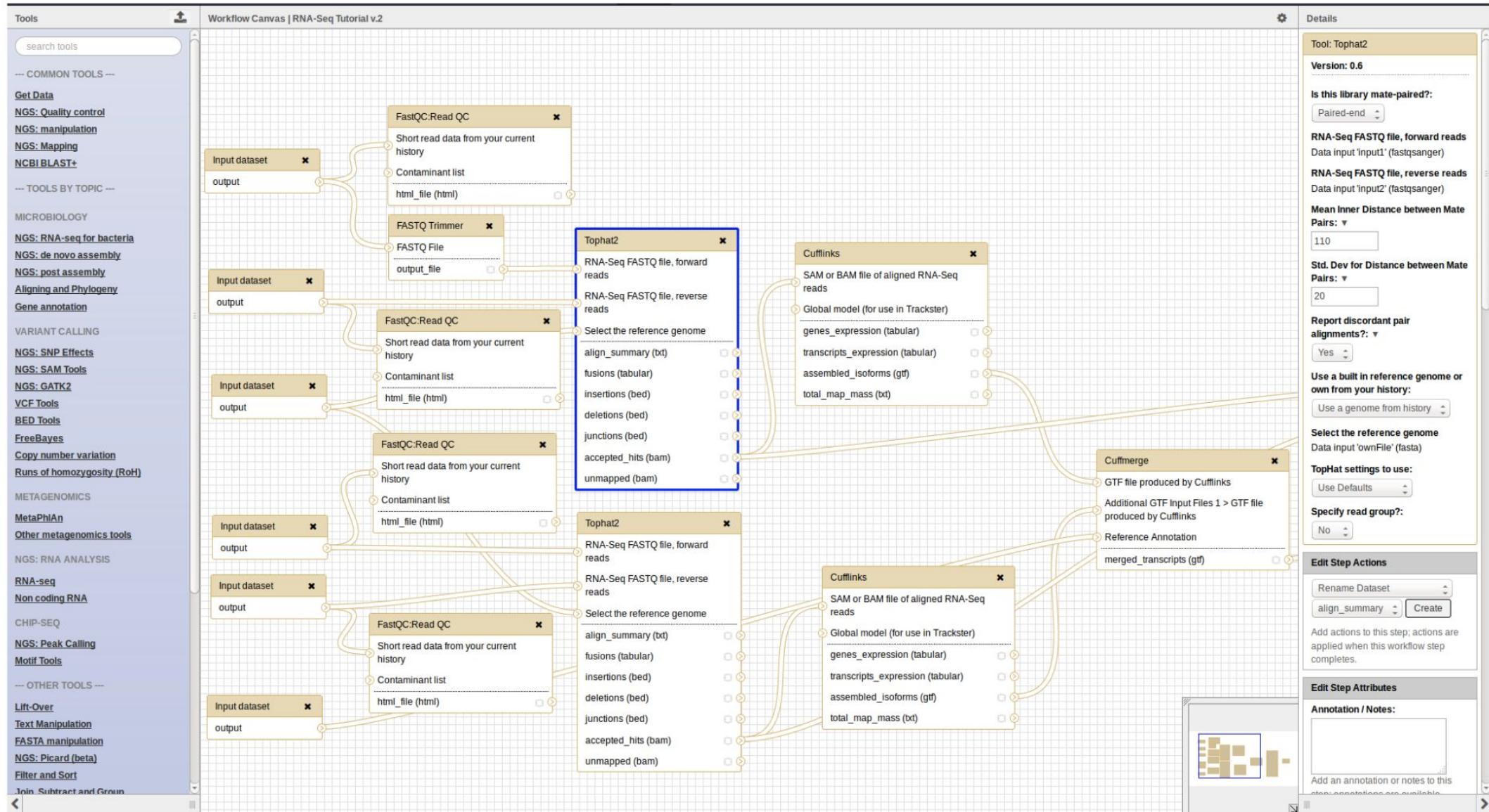
- Genome build specifies which genome assembly a dataset is associated with
 - e.g. mm10, hg38...
- Can be assigned by a tool or by the user
- Users can create custom genome builds
- New builds can be added by the admin

Database/Build

Build
Mouse July 2007 (NCBI37/mm9) (mm9)
Burmese python Sep. 2013 (Python_molurus_bivittatus-5.0.2/pytBiv1) (pytBiv1)
Burton's mouthbreeder Oct 2011 (AstBur1.0/hapBur1) (hapBur1)
Bushbaby Mar. 2011 (Broad/otoGar3) (otoGar3)
Bushbaby Dec. 2006 (Broad/otoGar1) (otoGar1)
C. angaria Oct. 2010 (WS225/caeAng1) (caeAng1)
C. brenneri Nov. 2010 (C. brenneri 6.0.1b/caePb3) (caePb3)
C. brenneri Feb. 2008 (WUGSC 6.0.1/caePb2) (caePb2)
C. brenneri Jan. 2007 (WUGSC 4.0/caePb1) (caePb1)

Workflow

Workflow Editor



Advantages of creating and saving workflows



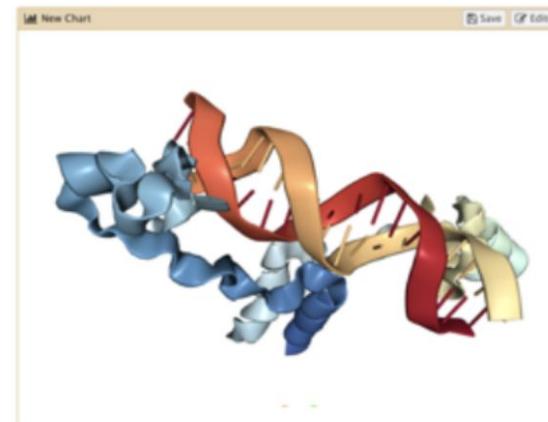
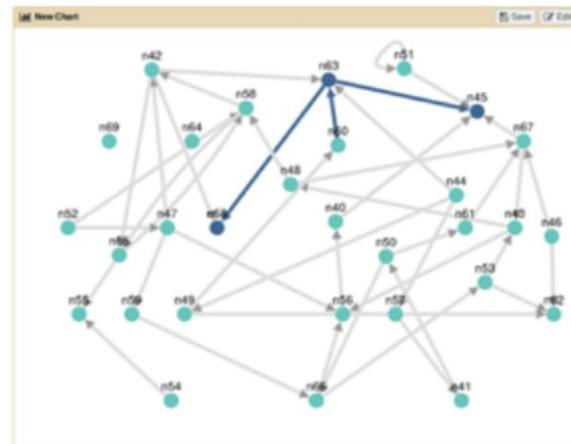
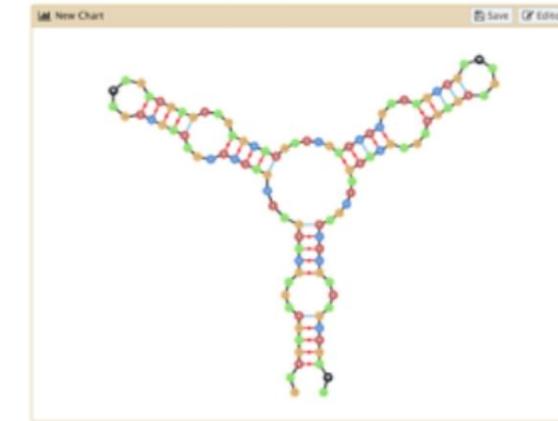
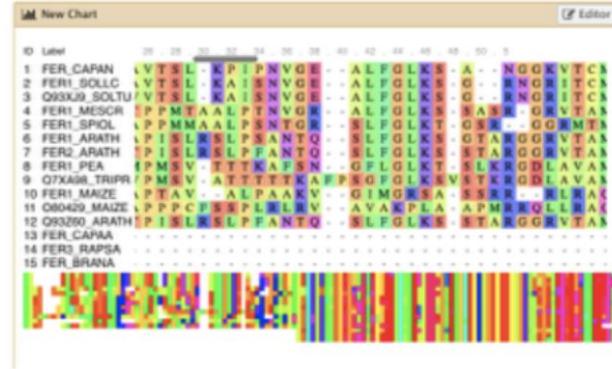
- **Re-run** the same analysis on different input data sets
- **Change parameters** before re-running a similar analysis
- Make use of the workflow job **scheduling**
 - jobs are submitted as soon as their inputs are ready
- Create **sub-workflows**: a workflow inside another workflow
- **Share** workflows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

Visualization in Galaxy

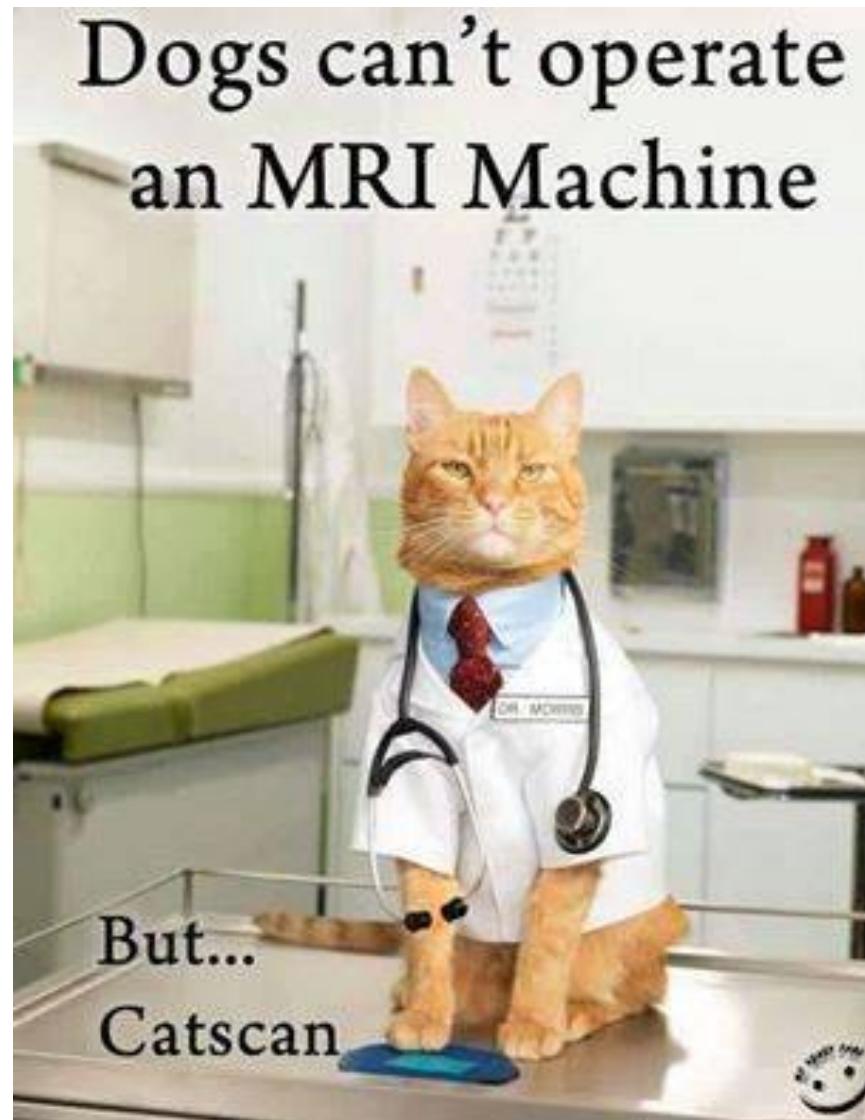


Karolinska
Institutet

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/>)