Session X – Galaxy

Miguel Juliá

BU-ISCIII

Unidades Comunes Científico Técnicas – SGSAFI-ISCIII

04-15 Noviembre 2019, 2ª Edición
Programa Formación Continua, ISCIII

Index

Galaxy:

- Computing in Biosciences
- Change of Paradigm
- What is Galaxy
- Workflows
- The porject
- Galaxy training

Computing in Biosciences

- Web-based platforms (i.e. Galaxy) and remote HPC

Pros	Cons
No need to storage intermediate files	Your data is in someone else's computer No backups or data management schemes
No need to install software Partial control over installed software	No control over installed software, versions and future availability
Graphic interface	No control over hidden parameters
Analysis are partially reproducible	Quotas

Change of Paradigm I

1 sample	multiple samples
Research only: NGS was still a new thing, no applications 10 years ago	**Many applications:** research, clinical, industrial, forensic, military, …
Reproducibility is not needed: Why would anyone reanalyse this?	**Reproducibility, scalability , portability and standardisation are required**
Storage is not an issue: files of 1 sample fits everywhere in my HDD, maybe I will copy it in a CD-ROM	**Storage is challenging:** storage, indexation and backup required, privacy and legal standards
Computing is simple: no need to worry about resources or optimisation	**Computing requires optimisation and lots of resources**

Change of Paradigm II

- Nowadays scientific computing paradigm

Pros	Cons
Data remains private Backups and data management schemes	High storage space Dedicated file systems Databases to index files
Control over software installed versions, open source programs	Many versions of the same software coexists
All parameters are available for the command	You have to understand all software variations
Analysis are reproducible and public	You have to publish and document your work

What is Galaxy I

Galaxy

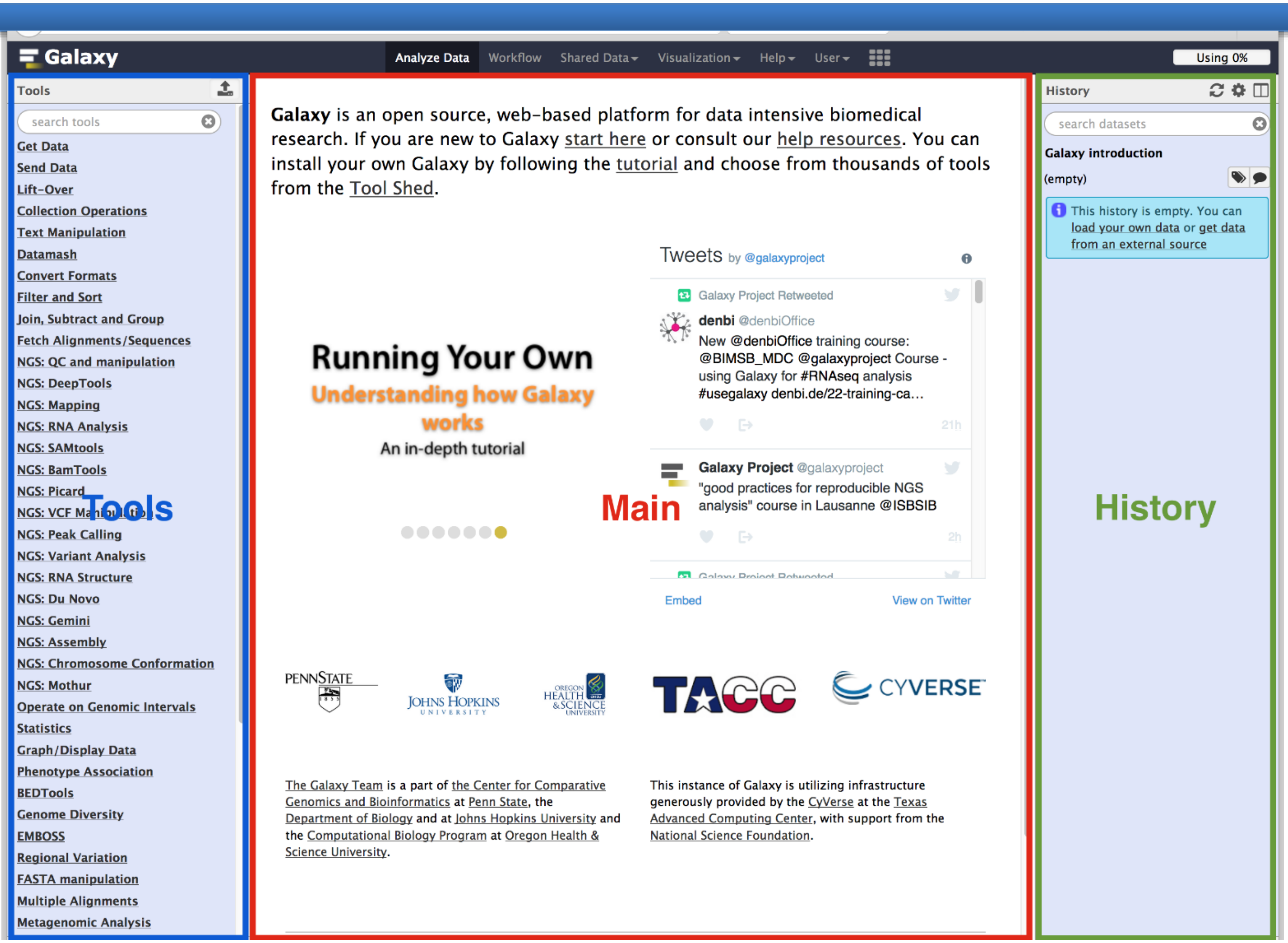
Data Intensive *analysis* for everyone

- 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

What is Galaxy II

- **Accessibility**
  - Users without programming experience can easily upload/retrieve data, run complex tools and workflows, and visualize data
- **Reproducibility**
  - Galaxy captures information so that any user can understand and repeat a complete computational analysis
- **Transparency**
  - Users can share or publish their analyses (histories, workflows, visualizations)
  - Pages: online Methods for your paper

What is Galaxy III



Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

This history is empty. You can load your own data or get data from an external source

Galaxy Project Retweeted
denbi @denbiOffice
New @denbiOffice training course: @BIMSB_MDC @galaxyproject Course - using Galaxy for #RNAseq analysis #usegalaxy denbi.de/22-training-ca…
21h

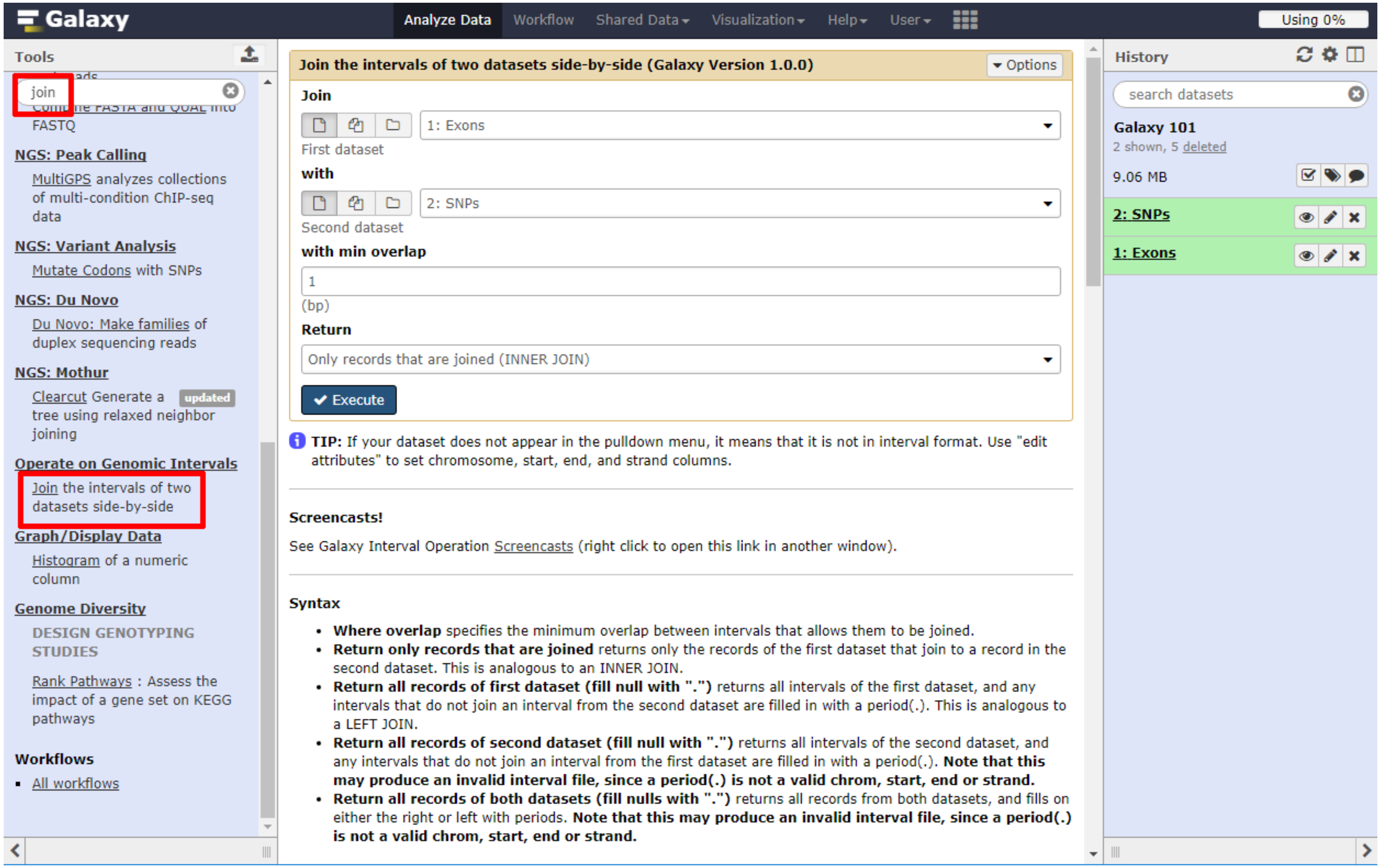
Galaxy Project @galaxyproject
"good practices for reproducible NGS analysis" course in Lausanne @ISBSIB
2h

Galaxy Project Retweeted

The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State, the Department of Biology and at Johns Hopkins University and the Computational Biology Program at Oregon Health & Science University.

This instance of Galaxy is utilizing infrastructure generously provided by the CyVerse at the Texas Advanced Computing Center, with support from the National Science Foundation.

What is Galaxy IV



join

Combine FASTA and QUAL into 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).

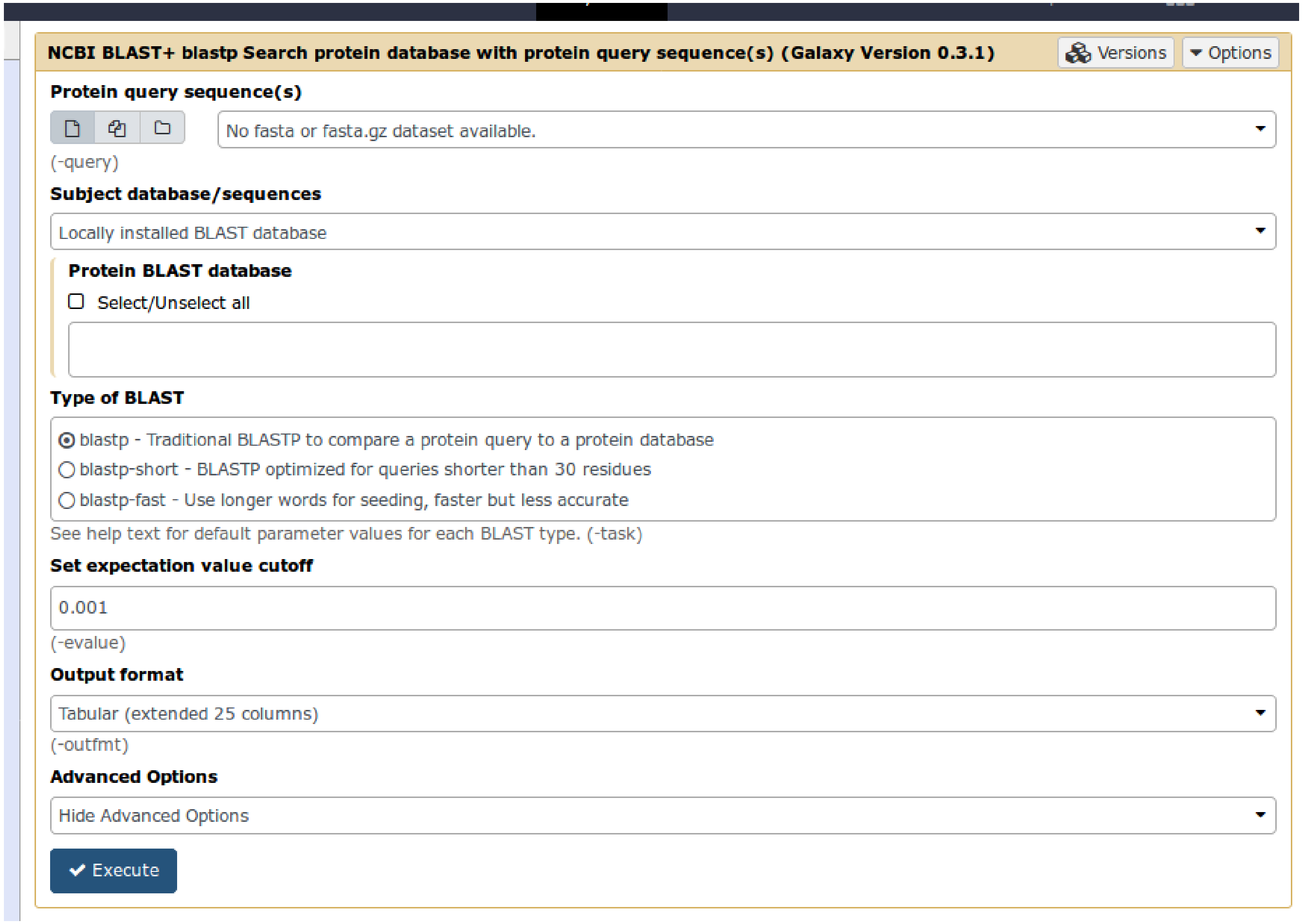
Syntax

- **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.**

Galaxy 101
2 shown, 5 deleted
9.06 MB

2: SNPs
1: Exons

What is Galaxy V



NCBI BLAST+ blastp Search protein database with protein query sequence(s) (Galaxy Version 0.3.1)

Protein query sequence(s)

No fasta or fasta.gz dataset available.

(-query)

Subject database/sequences

Locally installed BLAST database

Protein BLAST database

□ Select/Unselect all

Type of BLAST

● blastp - Traditional BLASTP to compare a protein query to a protein database
○ blastp-short - BLASTP optimized for queries shorter than 30 residues
○ blastp-fast - Use longer words for seeding, faster but less accurate

See help text for default parameter values for each BLAST type. (-task)

Set expectation value cutoff

0.001

(-evalue)

Output format

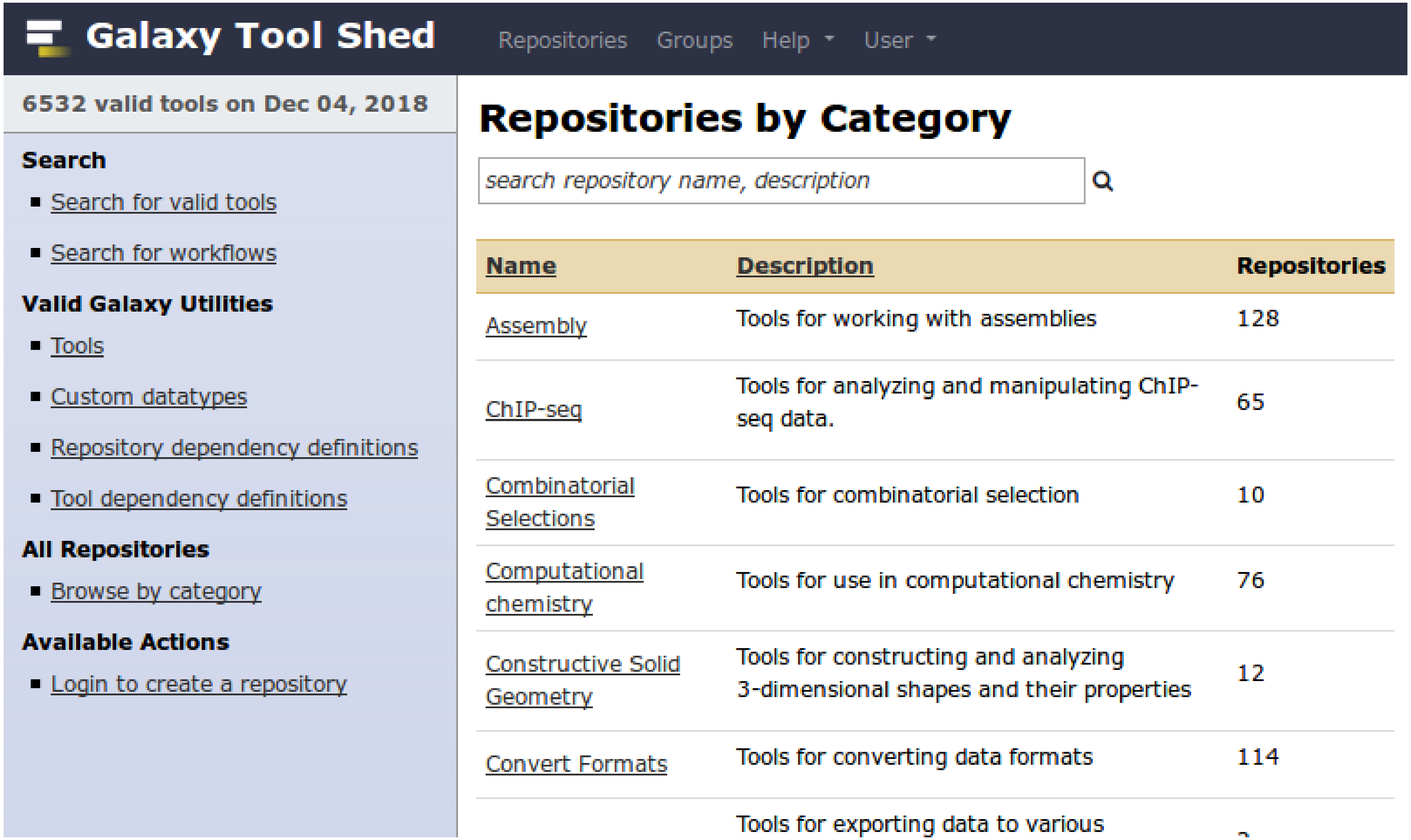
Tabular (extended 25 columns)

(-outfmt)

Advanced Options

Hide Advanced Options

What is Galaxy VI



6532 valid tools on Dec 04, 2018

Search

- Search for valid tools
- Search for workflows

Valid Galaxy Utilities

- Tools
- Custom datatypes
- Repository dependency definitions
- Tool dependency definitions

All Repositories

- Browse by category

Available Actions

- Login to create a repository

Repositories by Category

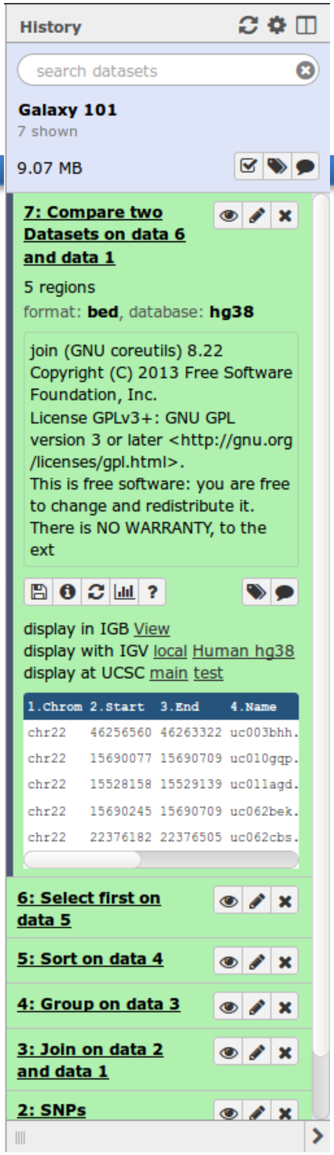
search repository name, description

Name	Description	Repositories
Assembly	Tools for working with assemblies	128
ChIP-seq	Tools for analyzing and manipulating ChIP-seq data.	65
Combinatorial Selections	Tools for combinatorial selection	10
Computational chemistry	Tools for use in computational chemistry	76
Constructive Solid Geometry	Tools for constructing and analyzing 3-dimensional shapes and their properties	12
Convert Formats	Tools for converting data formats	114
	Tools for exporting data to various	

What is Galaxy VII

- 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



History

search datasets

Galaxy 101

7 shown

9.07 MB

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

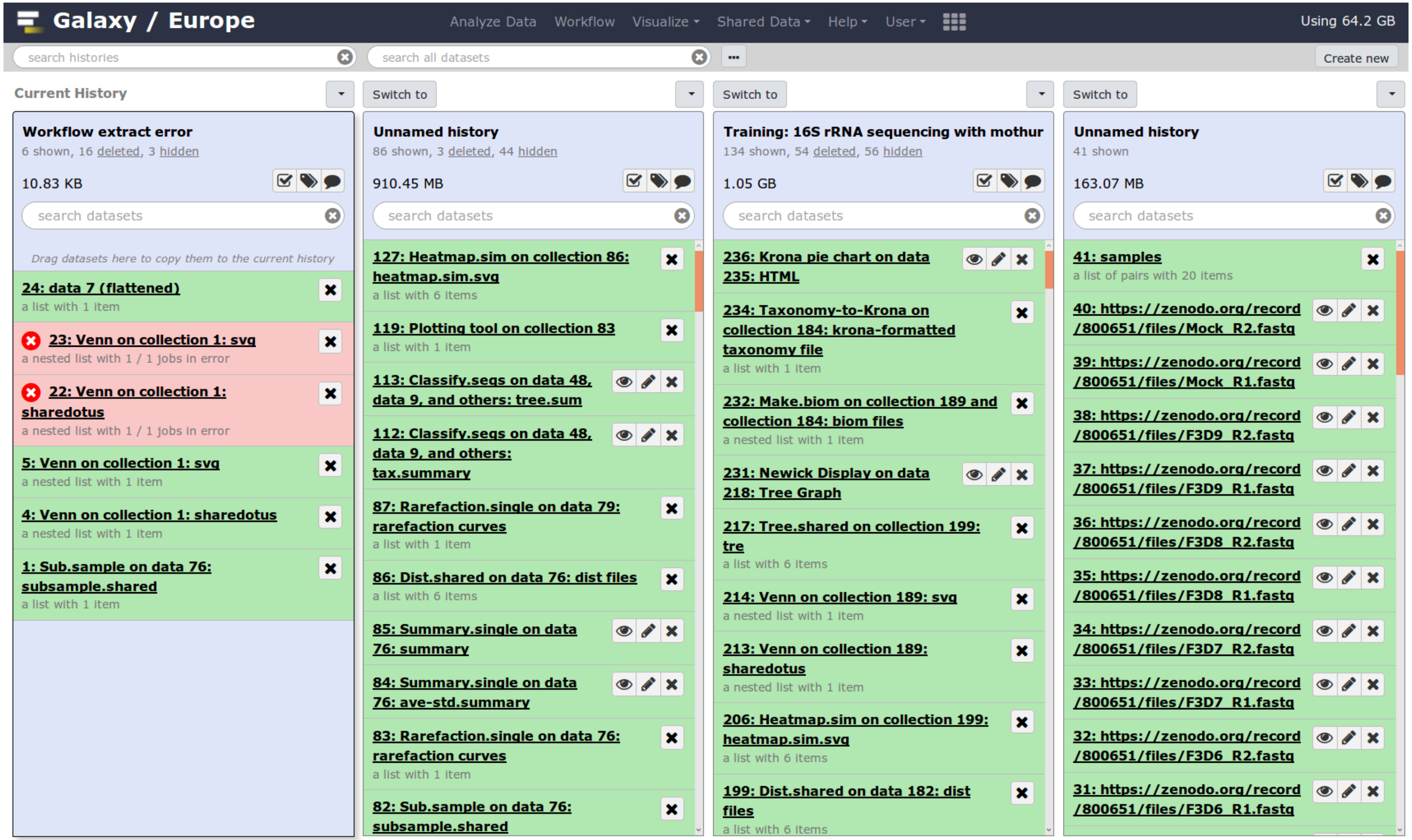
5: Sort on data 4

4: Group on data 3

3: Join on data 2 and data 1

2: SNPs

What is Galaxy VIII



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

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

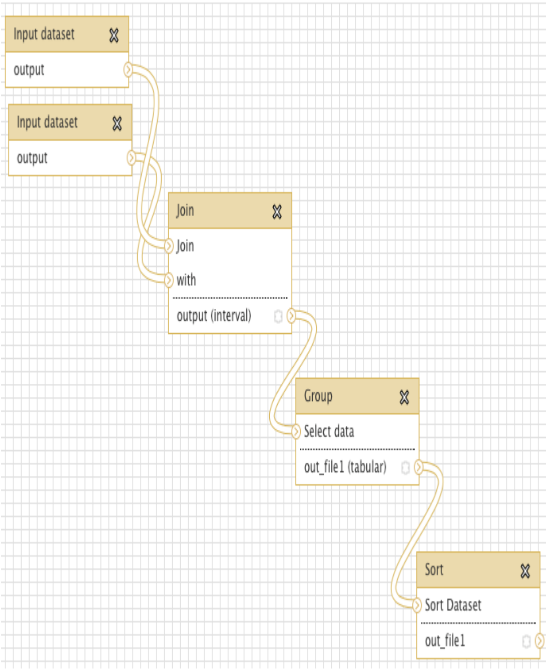
Unnamed history
41 shown
163.07 MB
search datasets
- 41: samples — a list of pairs with 20 items
- 40: https://zenodo.org/record/800651/files/Mock_R2.fastq
- 39: https://zenodo.org/record/800651/files/Mock_R1.fastq
- 38: https://zenodo.org/record/800651/files/F3D9_R2.fastq
- 37: https://zenodo.org/record/800651/files/F3D9_R1.fastq
- 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

Workflows I

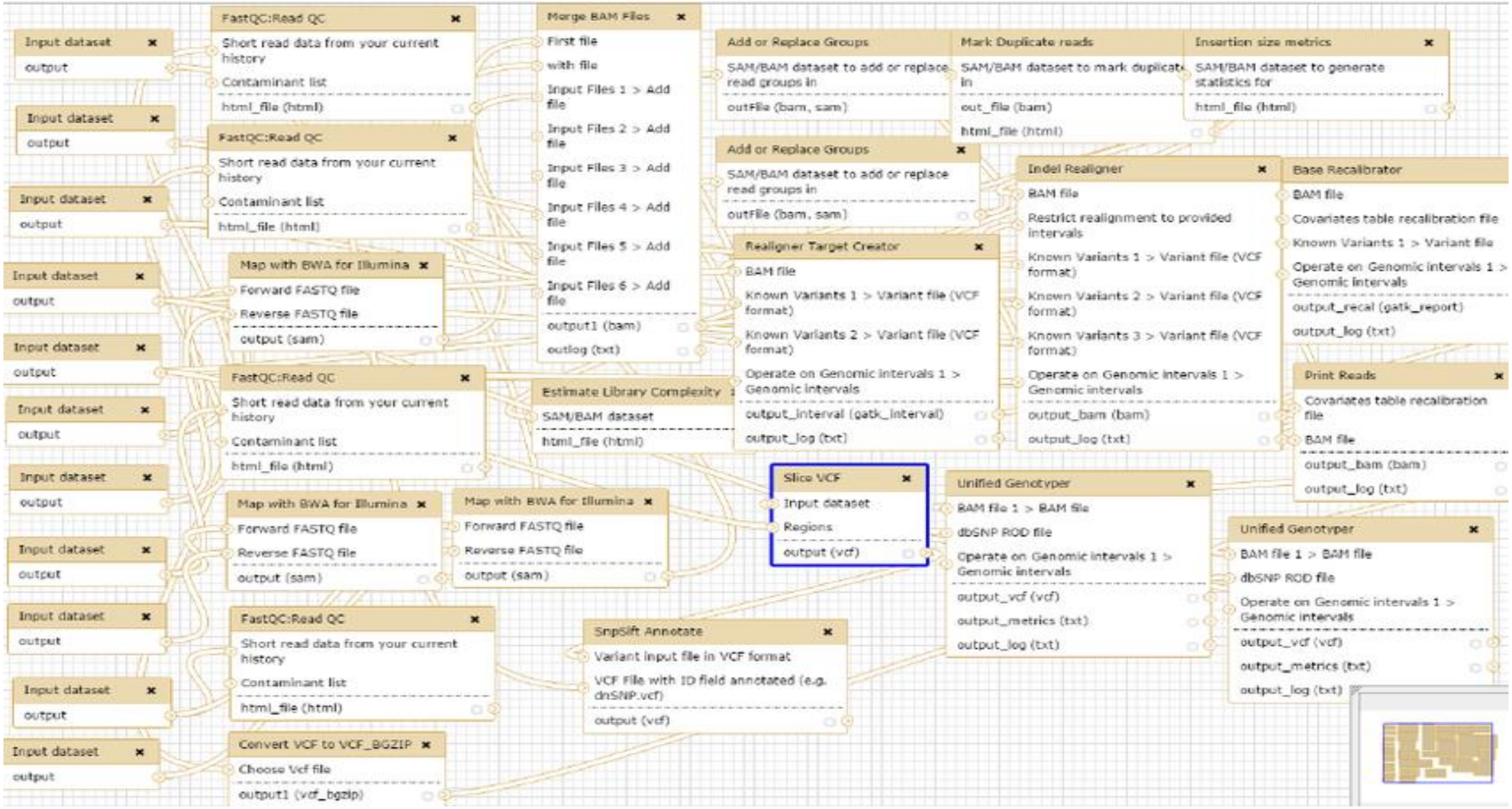
- Bioinformatic analyses invariably involve shepherding files through a series of transformations, called a **pipeline** or a **workflow.**

- These transformations are done by executable **command line software** written for Unix-compatible operating systems.

- They need to be **reproducible, easy to maintain, portable and scalable.**



Workflows II



Input dataset ✖
output

Input dataset ✖
output

Input dataset ✖
output

Input dataset ✖
output

Input dataset ✖
output

Input dataset ✖
output

Input dataset ✖
output

Input dataset ✖
output

Input dataset ✖
output

Input dataset ✖
output

Input dataset ✖
output

FastQC:Read QC ✖
Short read data from your current history
Contaminant list
html_file (html)

FastQC:Read QC ✖
Short read data from your current history
Contaminant list
html_file (html)

Map with BWA for Illumina ✖
Forward FASTQ file
Reverse FASTQ file
output (sam)

FastQC:Read QC ✖
Short read data from your current history
Contaminant list
html_file (html)

Map with BWA for Illumina ✖
Forward FASTQ file
Reverse FASTQ file
output (sam)

FastQC:Read QC ✖
Short read data from your current history
Contaminant list
html_file (html)

Convert VCF to VCF_BGZIP ✖
Choose Vcf file
output1 (vcf_bgzip)

Merge BAM Files ✖
First file
with file
Input Files 1 > Add file
Input Files 2 > Add file
Input Files 3 > Add file
Input Files 4 > Add file
Input Files 5 > Add file
Input Files 6 > Add file
output1 (bam)
outlog (txt)

Estimate Library Complexity
SAM/BAM dataset
html_file (html)

Map with BWA for Illumina ✖
Forward FASTQ file
Reverse FASTQ file
output (sam)

SnpSift Annotate ✖
Variant input file in VCF format
VCF File with ID field annotated (e.g. dnSNP.vcf)
output (vcf)

Add or Replace Groups
SAM/BAM dataset to add or replace read groups in
outFile (bam, sam)

Add or Replace Groups ✖
SAM/BAM dataset to add or replace read groups in
outFile (bam, sam)

Realigner Target Creator ✖
BAM file
Known Variants 1 > Variant file (VCF format)
Known Variants 2 > Variant file (VCF format)
Operate on Genomic intervals 1 > Genomic intervals
output_interval (gatk_interval)
output_log (txt)

Slice VCF ✖
Input dataset
Regions
output (vcf)

Mark Duplicate reads
SAM/BAM dataset to mark duplicat in
out_file (bam)
html_file (html)

Indel Realigner ✖
BAM file
Restrict realignment to provided intervals
Known Variants 1 > Variant file (VCF format)
Known Variants 2 > Variant file (VCF format)
Known Variants 3 > Variant file (VCF format)
Operate on Genomic intervals 1 > Genomic intervals
output_bam (bam)
output_log (txt)

Unified Genotyper ✖
BAM file 1 > BAM file
dbSNP ROD file
Operate on Genomic intervals 1 > Genomic intervals
output_vcf (vcf)
output_metrics (txt)
output_log (txt)

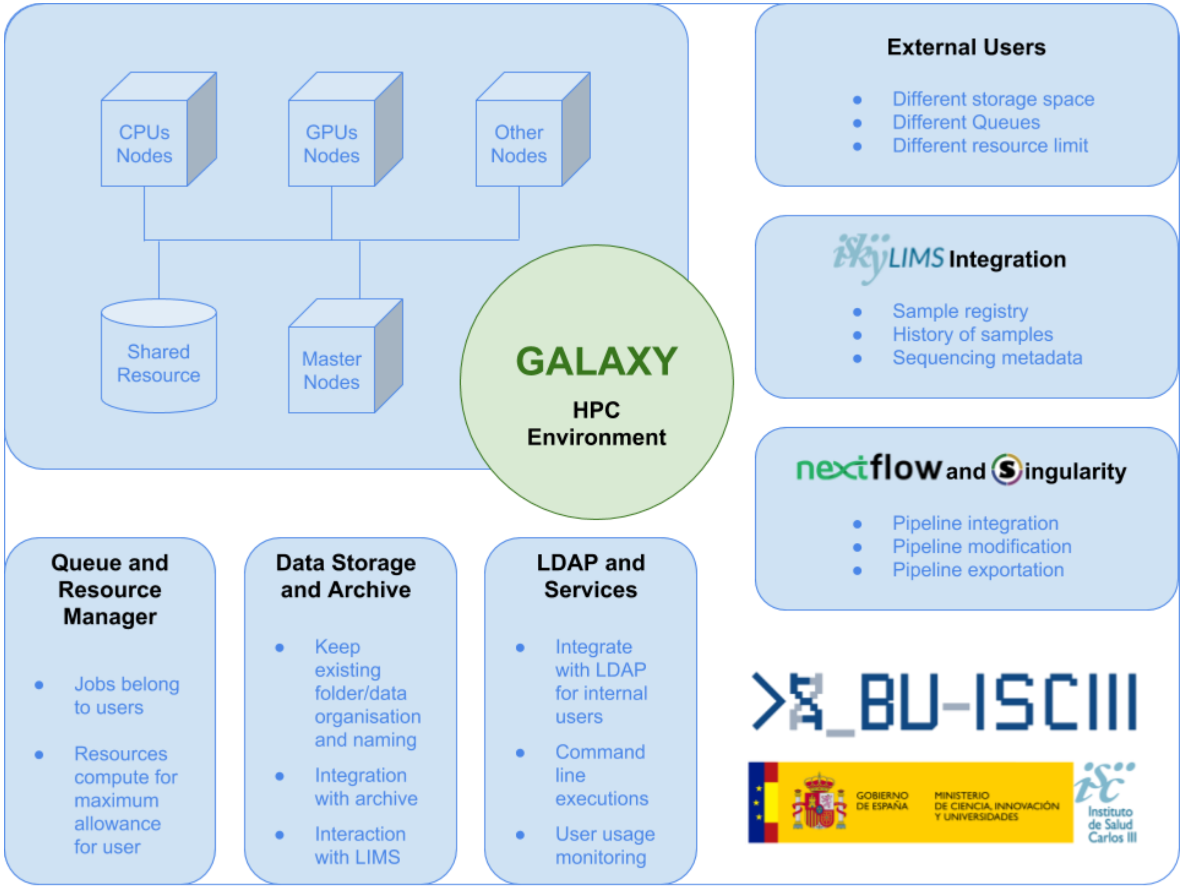
Insertion size metrics ✖
SAM/BAM dataset to generate statistics for
html_file (html)

Base Recalibrator
BAM file
Covariates table recalibration file
Known Variants 1 > Variant file
Operate on Genomic intervals 1 > Genomic intervals
output_recal (gatk_report)
output_log (txt)

Print Reads ✖
Covariates table recalibration file
BAM file
output_bam (bam)
output_log (txt)

Unified Genotyper ✖
BAM file 1 > BAM file
dbSNP ROD file
Operate on Genomic intervals 1 > Genomic intervals
output_vcf (vcf)
output_metrics (txt)
output_log (txt)

The Project



- Clinical data storage
- Hospitals
- Patient oriented research
- Training

Thanks for your attention!

And this is only the tip of the iceberg...
Check this if you wanna know what’s really going under the hood:

https://github.com/BU-ISCIII