



Resource for bioinformatics analyses and more

Felix HOMA INRA
Jean François DUFAYARD CIRAD

Bioinformatic in biologist teams



Problem including bioinformatic questions



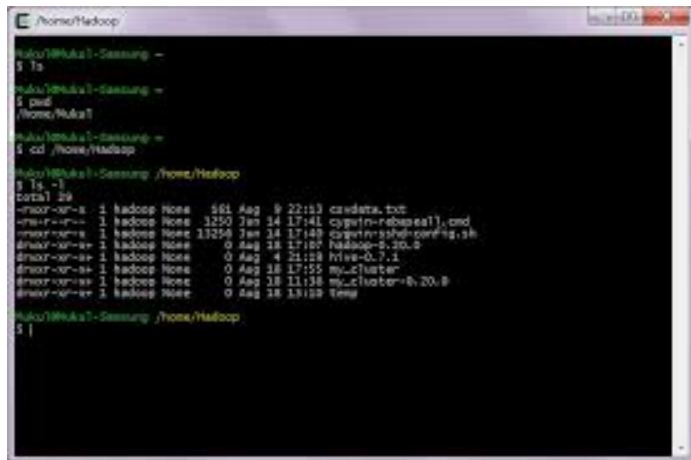
- Hard to find the right interlocutor
- Reaction times are variable



- He isn't omniscient
- He isn't omnipotent
- His time is limited

The "Do it yourself" procedure

Sometimes the analysis are "simple" and consists in using classical tools



```
huku@huku1:~$ cd /home/Hadoop
huku@huku1:~/Hadoop$ ls -l
total 28
-rw-r--r-- 1 hadoop None 561 Aug 9 22:03 czedata.txt
-rw-r--r-- 1 hadoop None 1250 Jan 14 17:41 cypwin-rebaseall.cmd
-rw-r--r-- 1 hadoop None 11256 Jan 14 17:40 cypwin-rebase-config.sh
-rw-r--r-- 1 hadoop None 0 Aug 15 17:05 hadoop-0.20.0
-rw-r--r-- 1 hadoop None 0 Aug 4 21:08 hive-0.7.1
-rw-r--r-- 1 hadoop None 0 Aug 18 17:55 my_cluster
-rw-r--r-- 1 hadoop None 0 Aug 18 11:38 my_cluster@h.20.0
-rw-r--r-- 1 hadoop None 0 Aug 18 13:03 Time
```



- Learning basic Unix commands
- Learning cluster specific commands
- Assume the technical and scientific survey
- Deal with file formats
- Deal with data transfers
- Scripting

Galaxy

Galaxy deals with some difficulties of "do it yourself"



Communication



Communication

 Galaxy

Galaxy is Free web service integrating a wealth of tools, compute resources, a great volume of data. It decrease the barrier to entry into data analysis by improving accessibility of the Galaxy platform.

Galaxy

- **Accessibility**
 - User friendly
 - Every tool under the same interface
 - No programming skill is needed
- **Reproductibility**
 - Allows to remake analysis and obtain expected result
- **Transparency**
 - Analysis can be documented and published
- **Integrative**
 - Includes many tools (computer science, statistics, text manipulation, visualization...)
- **Widely used** : [galaxy project](#) [Screencasts](#)



Objectives

- Objectives of training:
 - To have a first **overview** of Galaxy user interface
 - To be able to **manage data** in Galaxy
 - To be able to **create** and **use Workflows**
 - To be able to **share galaxy items**
 - **This session is not about SPECIFIC ANALYSIS**

Galaxy : Global view

Navigation menu

Toolbox

Current history

Analyze Data Workflow Shared Data Visualization Help User

Tools

Meta-genomic analyses

- Fetch taxonomic representation
- Summarize taxonomy
- Draw phylogeny
- Find diagnostic hits
- Find lowest diagnostic rank
- Poisson two-sample test

FASTA manipulation

- Compute sequence length
- Filter sequences by length
- Concatenate FASTA alignment by species
- FASTA-to-Tabular converter
- Tabular-to-FASTA converts tabular file to FASTA format
- FASTA Width formatter
- RNA/DNA converter
- Collapse sequences

NGS: QC and manipulation

- FASTQC: FASTQ/SAM/BAM
- Fastqc: Fastqc QC using FastQC from Bahraham



South Green[®]
bioinformatics platform

Welcome to GALAXY

Display panel
(Tool settings,
View file content)

be addressed to admin.bioinfo@cirad.fr

The GALAXY project is supported in part by NSF NHGRI,
and the Huck Institutes of the Life Sciences.

0 bytes

1:
Galaxy2-[IqB6 GCCAAT L003 R2.f
a/tq1.fastqsanger
empty
format: txt, database: ?
info: The uploaded file is empty

Galaxy : First run

Input files must be present in the current history

The screenshot shows the Galaxy web interface with several annotations:

- 1**: A blue box with the number '1' is placed over the 'Analyze Data' tab in the top navigation bar.
- 2**: A blue box with the number '2' and the text 'Upload dataset' is placed over the 'Upload dataset' button in the top right corner.
- 3**: A blue box with the number '3' and the text 'Select tool' is placed over the 'Tools' sidebar on the left.
- 4**: A blue box with the number '4' and the text 'Set parameters' is placed over the tool configuration form.
- 5**: A blue box with the number '5' and the text 'Execute' is placed over the 'Execute' button.

The tool configuration form for 'FASTQ to FASTA (version 1.0.0)' includes the following fields:

- FASTQ Library to convert:** 1: input1.fastq
- Discard sequences with no alignments:** yes
- Rename sequence names in output file (reduces file size):** yes

The 'Execute' button is highlighted with a blue box labeled '5 Execute'.

The 'History' panel on the right shows the following items:

- felix test (7.7 MB)
- 4: FASTQ to FASTA on data 1
- 3: reference.fasta
- 2: input2.fastq
- 1: input1.fastq

The 'Example' section shows the following text:

```
The following data in Solexa-FASTQ format:  
@CSHL_4_FC042GAMMII_2_1_517_596  
GGTCAATGATGAGTTGGCACTGTAGGCACCATCAAT  
+CSHL_4_FC042GAMMII_2_1_517_596  
40 40 40 40 40 40 40 40 40 38 40 40 40 40 14 40 40 40 40 36 40 13 14 24 24 9 24 9 40 10 10 15 40  
Will be converted to FASTA (with 'rename sequence names' = NO):  
>CSHL_4_FC042GAMMII_2_1_517_596  
GGTCAATGATGAGTTGGCACTGTAGGCACCATCAAT  
Will be converted to FASTA (with 'rename sequence names' = YES):  
>1  
GGTCAATGATGAGTTGGCACTGTAGGCACCATCAAT
```

This tool is based on [FASTX-toolkit](#) by Assaf Gordon.

Galaxy : Upload a dataset

The screenshot shows the Galaxy 'Upload File (version 1.1.3)' tool interface. On the left is a 'Tools' sidebar with a search bar and a list of tools. The 'Get Data' section is circled in blue. The main tool area has several sections: 'File Format' (set to 'Auto-detect'), 'File' (with a 'Parcourir...' button and a blue callout box 'Import data from your computer'), 'URL/Text' (with a text area and a blue callout box listing 'Web links (URLs)' and 'Copy / paste'), 'Files uploaded via FTP' (with a table showing no files), 'Convert spaces to tabs' (with a 'Yes' checkbox), and 'Genome' (with a dropdown menu). At the bottom is an 'Execute' button. On the right is a 'History' panel showing a job named 'felix test' (7.7 MB) with a list of four datasets: '4: FASTQ to FASTA on data 1', '3: reference.fasta', '2: input2.fastq', and '1: input1.fastq'. A blue callout box 'File format technical specification' points to the '1: input1.fastq' dataset. The 'FASTQ to FASTA' dataset shows a preview of FASTQ data.

Tools

search tools

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC microbes table browser
- BX table browser
- EBI SRA ENA SRA
- BioMart Central server
- CBI Rice Mart rice mart
- GrameneMart Central server
- modENCODE fly server
- Flymine server
- modENCODE modMine server
- Ratmine server
- YeastMine server
- metabolicMine server
- modENCODE worm server
- WormBase server
- EuPathDB server
- EncodeDB at NHGRI
- EpiGRAPH server
- HbVar Human Hemoglobin Variants and Thalassemias
- GenomeSpace import from file browser

Send Data

Upload File (version 1.1.3)

File Format:
Auto-detect
Which format? See help below

File:
Parcourir... Aucun fichier sélectionné

URL/Text:

- Web links (URLs)
- Copy / paste

Files uploaded via FTP:

File	Size	Date
Your FTP upload directory contains no files.		

Convert spaces to tabs:
 Yes
Use this option if you are entering intervals by hand.

Genome:

Execute

Auto-detect

The system will attempt to detect Axt, Fasta, Fastqsolexa, Gff, Gff3, Html, Lav, Maf, Tabular, Wiggle, Bed and Interval (Bed with headers) formats. If your file is not detected properly as one of the known formats, it most likely means that it has some format problems (e.g., different number of columns on different rows). You can still coerce the system to set your data to the format you think it should be. You can also upload compressed files, which will

History

felix test
7.7 MB

- 4: FASTQ to FASTA on data 1
- 3: reference.fasta
- 2: input2.fastq
- 1: input1.fastq

16,322 sequences
format: fastqsanger, database: dm3

uploaded fastq file

```
@RC10_HIJ51-E45454_0006:1:85:7371:16074#T46CTT  
GCAGS-TAGGSTCAACTGACAGATAGGATTAAGCTGCATGCATAGG  
+  
feeffffffffffffffedffc^eefffeccfefffffeccadfd  
@RC10_HIJ51-E45454_0006:1:54:10343:6310#T46CTT  
TGAAG-TTGAAGGAGGATGAGTGGCCGCACTCTCCCGGCTCAGATC
```

File format technical specification

Galaxy : Color code

Galaxy associate a color for each step of analysis

The screenshot shows the Galaxy web interface with a table of datasets. The table has columns for Datasets, Tags, Sharing, Size on Disk, Created, Last Updated, and Status. The datasets are color-coded based on their status:

- Green: Analysis well done
- Yellow: Pending analysis
- Red: Error

Blue callout boxes with red arrows point to these color-coded cells in the table. The 'Running analysis' box points to a yellow cell. The 'Analysis well done' box points to a green cell. The 'Error' box points to a red cell. The 'Pending analysis' box points to a yellow cell.

	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated	Status
Unnas history	13	2	2	7	0 Tags	65.2 MB	less than a minute ago
Unnas history		0 Tags		0 bytes	Dec 21, 2012	~ 23 hours ago	curr hist
arcad_workflow_big				3.7 GB	May 07, 2012	Dec 21, 2012	
gwas_test				39.9 MB	Dec 13, 2012	Dec 21, 2012	
galaxy_test	20			52.1 MB	May 04, 2012	Nov 13, 2012	
Unnas history				0 bytes	Oct 01, 2012	Oct 01, 2012	

History panel on the right shows a list of output files with their status:

- 79: Output file, snp information (green)
- 78: Output file, snp genotype MAF (red)
- 77: Output file, snp information (yellow)
- 76: Output file, snp genotype MAF (yellow)
- 75: Output file, snp information (green)

Galaxy : Current history/datasets

History annotation/tags

Datasets

Re-run analysis

Dataset metadata

Dataset actions

The screenshot shows the Galaxy interface's history panel. At the top, a header 'History' includes a refresh icon and a settings gear icon. Below this, a history entry 'felix test' is shown with a size of '7.7 MB' and icons for sharing and help. The main content area displays a workflow step '4: FASTQ to FASTA on data_1' in a green box, with sub-steps '3: reference.fasta' and '2: input2.fastq'. The step '4' includes a description '3 sequences', 'format: fasta, database: dm3', and a text input field 'uploaded fasta file'. Below the input field are icons for saving, sharing, and re-running the analysis. A scrollable area shows sequence data for 'LOC_0901g44110.1'. At the bottom, dataset '1: input1.fastq' is visible with icons for viewing, editing, and deleting.

The screenshot shows the 'Actions' menu for a history item. It is divided into two sections: 'HISTORY LISTS' and 'CURRENT HISTORY'. The 'HISTORY LISTS' section includes 'Saved Histories' and 'Histories Shared with Me'. The 'CURRENT HISTORY' section includes: 'Create New', 'Copy History', 'Copy Datasets', 'Share or Publish', 'Extract Workflow', 'Dataset Security', 'Resume Paused Jobs', 'Collapse Expanded Datasets', 'Include Deleted Datasets', 'Include Hidden Datasets', 'Unhide Hidden Datasets', 'Delete Hidden Datasets', 'Purge Deleted Datasets', 'Show Structure', 'Export to File', 'Delete', and 'Delete Permanently'.

Actions

Galaxy : Histories management

The screenshot shows the Galaxy Histories management interface. On the left, a sidebar lists 'HISTORY LISTS' with 'Saved Histories' circled in blue. Below it are sections for 'Histories Shared with Me' and 'CURRENT HISTORY' with various actions like 'Create New', 'Copy History', etc. The main area is titled 'Saved Histories' and contains a search bar and a table of history entries. A context menu is open over the 'felix' entry, showing actions like 'Switch', 'View', 'Share or Publish', 'Rename', 'Delete', and 'Delete Permanently'. A blue callout box points to this menu with the text 'Actions on one history'. Another blue callout box points to the bottom of the table with the text 'Multiple selection and action'. A third blue callout box points to the 'current history' label on the right side of the table. At the bottom, a control bar shows 'For 0 selected histories:' followed by buttons for 'Rename', 'Delete', 'Delete Permanently', and 'Undelete'.

Saved Histories

search history names and tags
Advanced Search

<input type="checkbox"/>	Name	Data	Disk	Created	Last Updated ↑	Status
<input type="checkbox"/>	CRN52	36	3	1 day ago	~ 1 hour ago	
<input type="checkbox"/>	galaxy tra	0 Tags	0 bytes	~ 5 hours ago	~ 2 hours ago	
<input type="checkbox"/>	felix	4 Tags	7.7 MB	1 day ago	~ 5 hours ago	current history
<input type="checkbox"/>	gau	17 4 Tags	99.4 MB	1 day ago	1 day ago	
<input type="checkbox"/>	CRN51	1 0 Tags	177.5 MB	1 day ago	1 day ago	
<input type="checkbox"/>			117.6 MB	1 day ago	1 day ago	
<input type="checkbox"/>			90.8 MB	1 day ago	1 day ago	
<input type="checkbox"/>			269.9 MB	1 day ago	1 day ago	
<input type="checkbox"/>	CRN24	1	686.7 MB	1 day ago	1 day ago	
<input type="checkbox"/>	CRN20	1 0 Tags	703.1 MB	1 day ago	1 day ago	

For 0 selected histories:

Galaxy : Best practices

- Never use any **accents** or **special characters** in any metadata of Galaxy : source of bugs
- Create new history for each new analysis
- Always give an **explicit name** to interesting histories/datasets : 2 histories/datasets can wear the same name
- **Delete** not helpful Galaxy items: think about community
- Always annotate with **tags** interesting histories and datasets : allow easy access to data, improve reproducibility of analysis

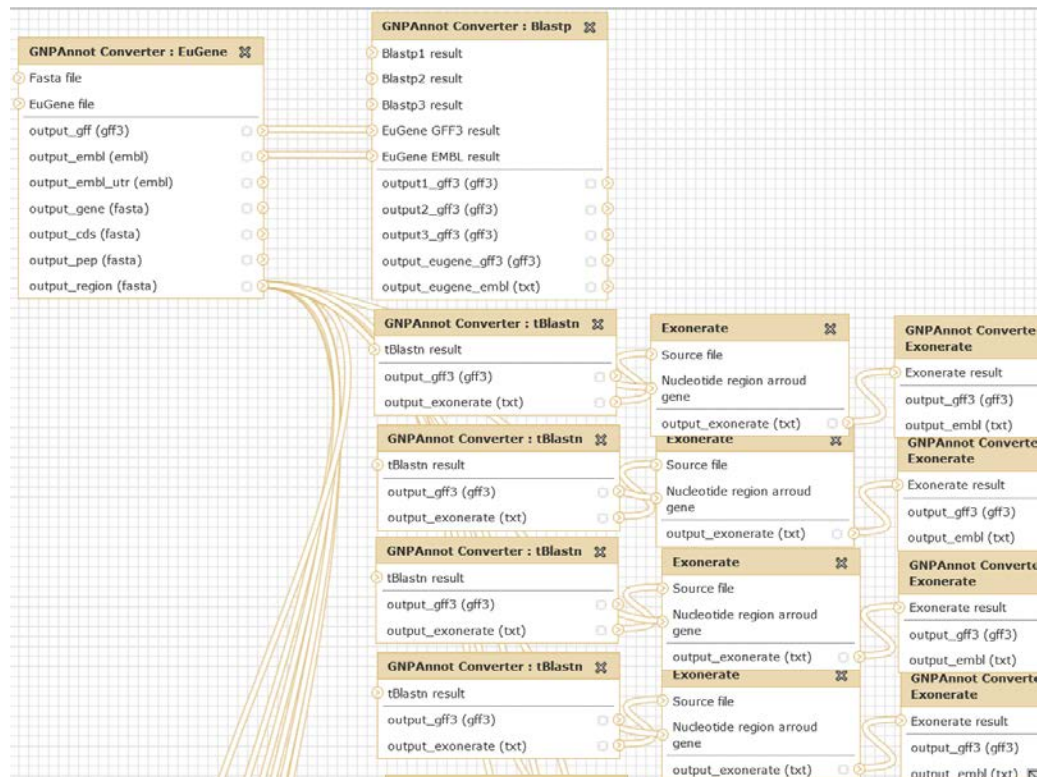
Practice



Practical session: Step 1
Changement de format
Création, renommage
d'historique

Workflows

- Means a simple or very complex series of operations, carried out automatically. Operations refer to analysis steps, data reformatting steps.



Workflow : build and configure

The image displays the Galaxy workflow editor interface. At the top, the navigation bar includes 'Analyze Data', 'Workflow' (highlighted with a red box), 'Shared Data', 'Visualization', 'Admin', 'Help', and 'User'. The main workspace is titled 'Workflow Canvas | Workflow ARCAD Step1 (imported from uploaded file)'. On the left, a sidebar lists various tool categories such as 'Get Data', 'Send Data', 'TOOLS', 'Convert Formats', 'Evolution', 'Filter and Sort', 'Fetch Sequences', 'Gene/Protein prediction', 'SniPlay', 'Population Analysis', 'ESTtik', 'SAT', 'NGS: Quality Control', 'NGS: Mapping', 'NGS: SAM/BAM Manipulations', 'NGS: SNP Detection', 'Protein Structures', 'Sequence comparisons', 'Genomics', 'BACCHUS Pipeline', 'UNTESTED TOOLS', 'ENCODE Tools', 'Lift-Over', 'Text Manipulation', 'Join, Subtract and Group', 'Extract Features', 'Fetch Alignments', and 'Get Genomic Scores'. The central 'Edition canevas' (highlighted with a blue box) shows a workflow diagram with several tool steps connected by arrows. The steps include 'Input dataset', 'FASTQ Groomer', 'File to groom', 'Concatenate datasets', 'Concatenate Dataset', 'Dataset 1 > Select', 'Reformat read identifier', 'FASTQ reads', 'FASTQ de-interlacer', 'FASTQ reads', 'FASTQ de-interlacer', 'FASTQ Groomer', 'Cutadapt', and 'Filter FASTQ'. On the right, the 'Parameters' panel (highlighted with a blue box) is open, showing configuration options for the 'File to groom' step. The parameters include 'Version: 1.0.4', 'File to groom' (Data input 'input_file' (fastq)), 'Input FASTQ quality scores type' (set to 'Sanger (recommended)'), 'Output FASTQ quality scores type' (set to 'Sanger (recommended)'), 'Force Quality Score encoding' (set to 'ASCII'), and 'Summarize input data' (set to 'Summarize Input'). Below these are sections for 'Edit Step Actions' (with 'Rename Dataset' and 'Create' buttons) and 'Edit Step Attributes' (with an 'Annotation / Notes' field).

Workflow : build and configure

The image shows a screenshot of the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', 'Help', and 'User'. The main area is titled 'Workflow Canvas | Workflow ARCAD Step1 (imported from uploaded file)'. On the left, there is a 'Tools' sidebar with a search bar and various tool categories like 'Get Data', 'Send Data', 'Convert Formats', 'Evolution', 'Filter and Sort', 'Fetch Sequences', 'Gene/Protein prediction', 'SNIPlay', 'Population Analysis', 'ESTtik', 'SAT', 'NGS: Quality Control', 'NGS: Mapping', 'NGS: SAM/BAM Manipulations', 'NGS: SNP Detection', 'Protein Structures', 'Sequence comparisons', 'Genomics', 'BACCHUS Pipeline', and 'UNTESTED TOOLS'. The workflow canvas itself contains several tool nodes connected by lines. A large blue callout box with a white exclamation mark is positioned over the 'FASTQ Groomer' tool. Another blue callout box points to the 'Concatenate datasets' tool. A third blue callout box points to the 'FASTQ de-interlacer' tool. A fourth blue callout box points to the 'Cutadapt' tool. The right side of the interface shows a 'Details' panel for the 'FASTQ Groomer' tool, including a 'Summarize input data' dropdown and 'Edit Step Actions' like 'Rename Dataset' and 'Create'. The bottom right corner has a 'What it does' section.

Galaxy

Analyze Data Workflow Shared Data Visualization Admin Help User

South Green Using 3.4 GB

Tools Options Workflow Canvas | Workflow ARCAD Step1 (imported from uploaded file) Options Details

search tools

Get Data

Send Data

TOOLS

Convert Formats

Evolution

Filter and Sort

Fetch Sequences

Gene/Protein prediction

SNIPlay

Population Analysis

ESTtik

SAT

NGS: Quality Control

NGS: Mapping

NGS: SAM/BAM Manipulations

NGS: SNP Detection

Protein Structures

Sequence comparisons

Genomics

BACCHUS Pipeline

UNTESTED TOOLS

ENCOD

Lift-Ov

Text M

Join, S

Extract

Fetch Alignments

Get Genomic Scores

Input dataset

output

FASTQ Groomer

File to groom

output_file (fastqsand fastqsolexa, fastqill

Concatenate datasets

Concatenate Dataset

Dataset 1 > Select

out_file1

Reformat read identifier

FASTQ reads

output_file

FASTQ de-interlacer

FASTQ reads

output1_pairs_file

output2_pairs_file

output1_singles_file

output2_singles_file

Summarize input data: Summarize Input

Edit Step Actions

Rename Dataset

output_file Create

This file will be displayed after analysis

This file will be hidden, but could be display afterwards

If nothing is checked, everything is considered checked

Cutadapt

file to trim

report (txt)

output

rest_output

too_short_output

untrimmed_output

Filter

output_file

What it does

Add an annotation or notes to this step; annotations are available when a workflow is viewed.

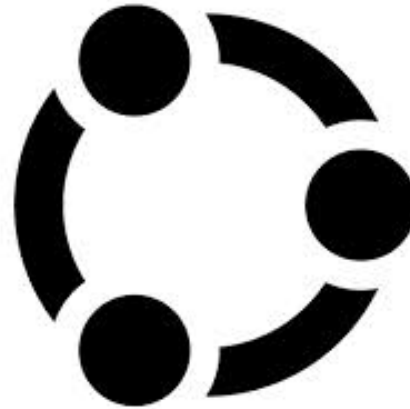
Practice



Practical session: Step 2

Shareable Items

- Galaxy items
 - History
 - Datasets
 - Workflow
 - Pages



Sharing with galaxy

- Share Galaxy items (3 ways)
 - **By link:** allowed users can make a local copy in their own workflow space, using provided link
 - **By publication:** every user can access to your workflow in the menu “Shared Data ->Published Workflows”
 - **By email:** share the workflow directly with a specific individual user

Sharing with galaxy

Saved Histories

[Advanced Search](#)

<input type="checkbox"/>	Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated	Status
<input type="checkbox"/>	CRN52	36	2	0 Tags	297.5 MB	1 day ago	~ 1 hour ago	
<input type="checkbox"/>	galaxy tra			0 Tags	0 bytes	~ 5 hours ago	~ 2 hours ago	
<input type="checkbox"/>	felib		4 Tags	7.7 MB	1 day ago	~ 5 hours ago	current history	
<input type="checkbox"/>	gal		4 Tags	99.4 MB	1 day ago	1 day ago		
<input type="checkbox"/>	CRN01	1		0 Tags	177.5 MB	1 day ago	1 day ago	
<input type="checkbox"/>	CRN49	1		0 Tags	117.6 MB	1 day ago	1 day ago	
<input type="checkbox"/>	CRN45	1		0 Tags	90.8 MB	1 day ago	1 day ago	
<input type="checkbox"/>	CRN31	1		0 Tags	269.9 MB	1 day ago	1 day ago	
<input type="checkbox"/>	CRN24	1		0 Tags	68			
<input type="checkbox"/>	CRN20	1		0 Tags	70			

For 0 selected histories:

- HISTORY LISTS
- Saved Histories
- Histories Shared with Me
- CURRENT HISTORY
- Create New
- Copy History
- Copy Datasets
- Share or Publish
- Extract Workflow
- Dataset Security
- Resume Paused Jobs
- Collapse Expanded Datasets
- Include Deleted Datasets
- Include Hidden Datasets
- Unhide Hidden Datasets
- Delete Hidden Datasets
- Purge Deleted Datasets
- Show Structure
- Export to File
- Delete
- Delete Permanently

Your workflows

Name

Workflow card with context menu:

- Edit
- Run
- Share or Publish
- Download or Export
- Copy
- Rename
- View
- Delete

Workflow shared with you by others

No workflow shared with you.

Other workflow cards visible:

- Workflow card with 'View' button
- Workflow card with 'Copy' button

Sharing with galaxy : access to shared data

The screenshot shows the Galaxy web interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The 'Shared Data' menu is open, showing options: 'Data Libraries', 'Published Histories', 'Published Workflows', 'Published Visualizations', and 'Published Pages'. A blue callout box on the left contains the text 'All published and shared items' and points to the 'Shared Data' menu. In the background, a table is partially visible with a header '# of Steps' and a value '9'. A 'Create new workflow' button is also present.

All published and shared items

of Steps
9

Create new workflow

Practice



Practical session: Step 3

More with Galaxy

- Security of data
 - Manage permission of shared data
- Use Galaxy pages to initiate scientist paper
- Install tools and new visualization
- Custom Galaxy look and feel
- Use Galaxy with the cloud
- Interaction Galaxy and homemade tools