Galaxy

An open, web-based platform for data intensive research.

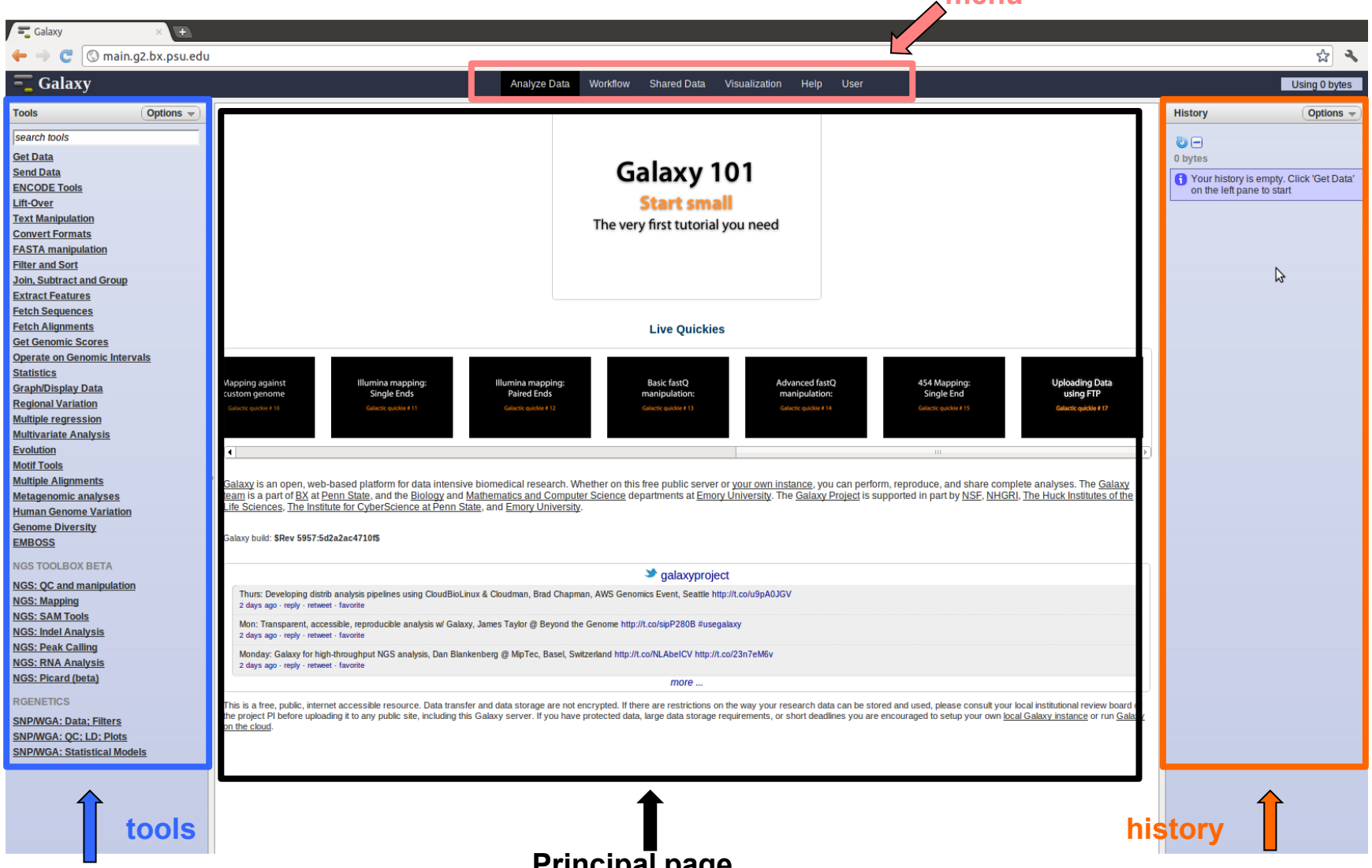
URGI-MIG, Galaxy, 3/26/2012
yluo@versailles.inra.fr

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

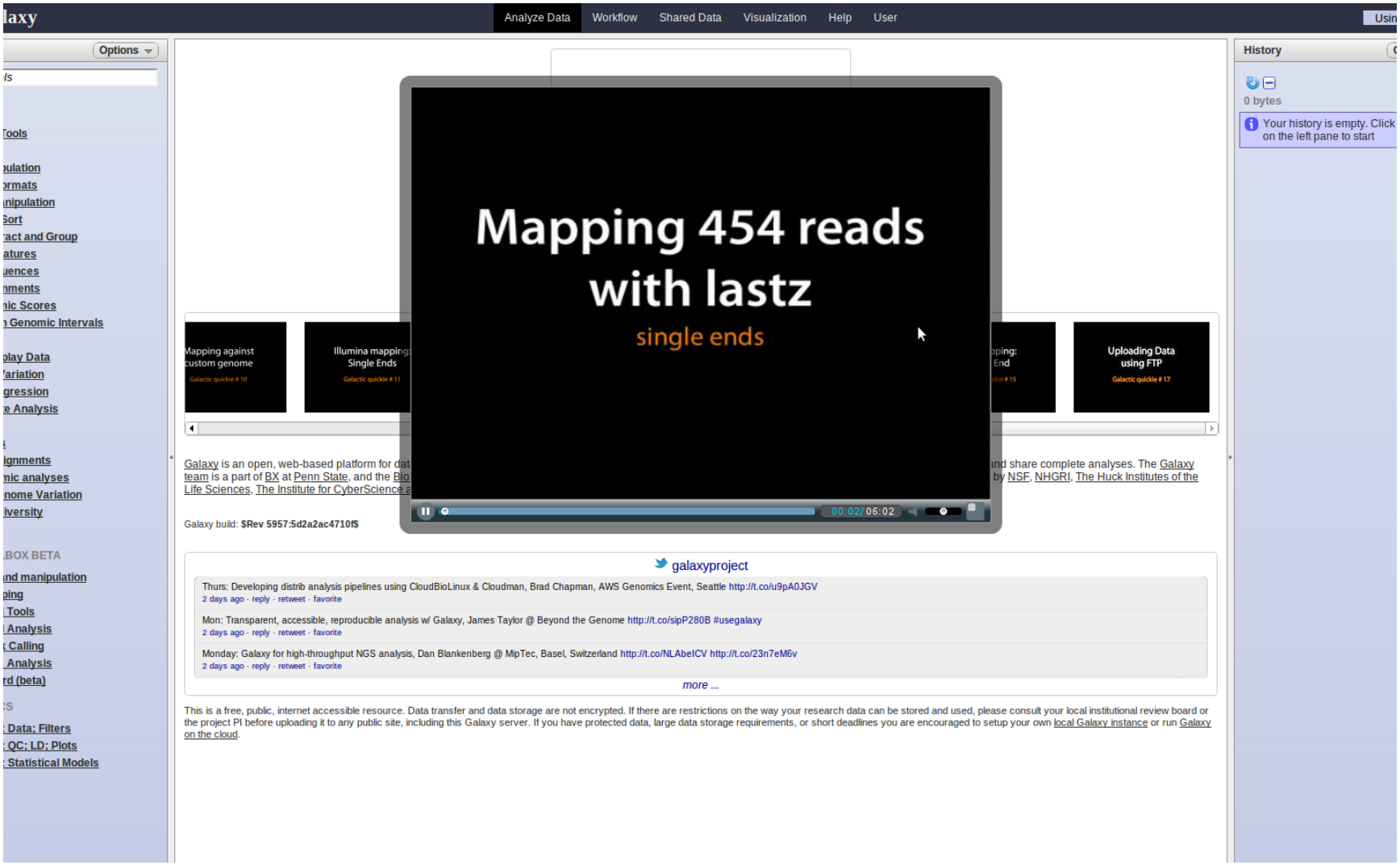
Galaxy

- Galaxy is a framework for integrating computational tools. More @ http://galaxy.psu.edu/
- It allows nearly any tool that can be run from the command line to be wrapped in a well-defined interface.
- On top of these tools, Galaxy provides an accessible environment for interactive analysis that transparently tracks the details of analysis steps, a workflow system for convenient reuse, data management, sharing, publishing, and so on.
- Two (free) ways to perform analysis with Galaxy:
  Using the open web server @ http://main.g2.bx.psu.edu/
  Install his own instance - Tuto @ http://wiki.g2.bx.psu.edu/Admin/Get Galaxy/
- Another way is to use the Amazon Galaxy cloud

Galaxy homepage



Galaxy


Analyze Data
Workflow
Shared Data
Visualization
Help
User
Options
History
0 bytes
Your history is empty. Click on the left pane to start
Mapping 454 reads with lastz
single ends
Uploading Data using FTP
Galaxy build: $Rev 5957:5d2a2ac4710f$
galaxyproject
Thurs: Developing distrib analysis pipelines using CloudBioLinux & Cloudman, Brad Chapman, AWS Genomics Event, Seattle http://t.co/u9pA0JGV
Mon: Transparent, accessible, reproducible analysis w/ Galaxy, James Taylor @ Beyond the Genome http://t.co/sipP280B #usegalaxy
Monday: Galaxy for high-throughput NGS analysis, Dan Blankenberg @ MipTec, Basel, Switzerland http://t.co/NLAbeICV http://t.co/23n7eM6v
more ...
This is a free, public, internet accessible resource. Data transfer and data storage are not encrypted. If there are restrictions on the way your research data can be stored and used, please consult your local institutional review board or the project PI before uploading it to any public site, including this Galaxy server. If you have protected data, large data storage requirements, or short deadlines you are encouraged to setup your own local Galaxy instance or run Galaxy on the cloud.

Galaxy

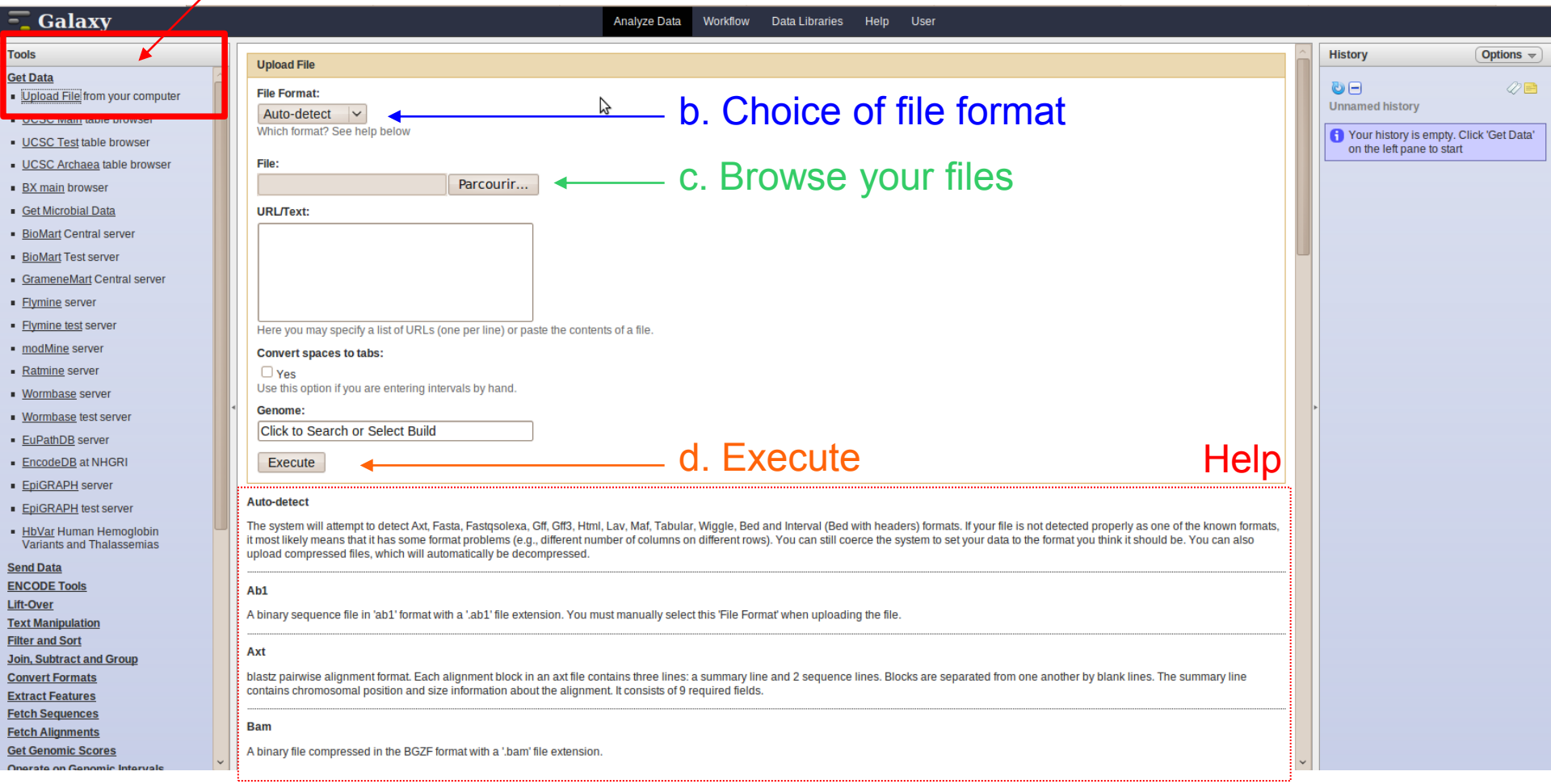
- **NGS (Illumina, Roche-454, AB-SOLiD)**
  - QC manipulation
  - mapping (BWA/Bowtie)
  - SAM tools
  - Indel analysis
  - RNA analysis
  - …
- **SNP**
  - Filters (Varscan)
  - QC, LD plots
  - statistical Models
- **General treatement data**
  - Get Data: upload file, BioMart, DB server, …
  - Text manipulation: add, merge, cut columns, …
  - Fasta manipulation: convert, compute fasta files, …
  - Statistic: summary, count, correlation, …
  - Graph
  - Alignments
  - Evolution
  - Metagenomics
  - …

Existing Tools

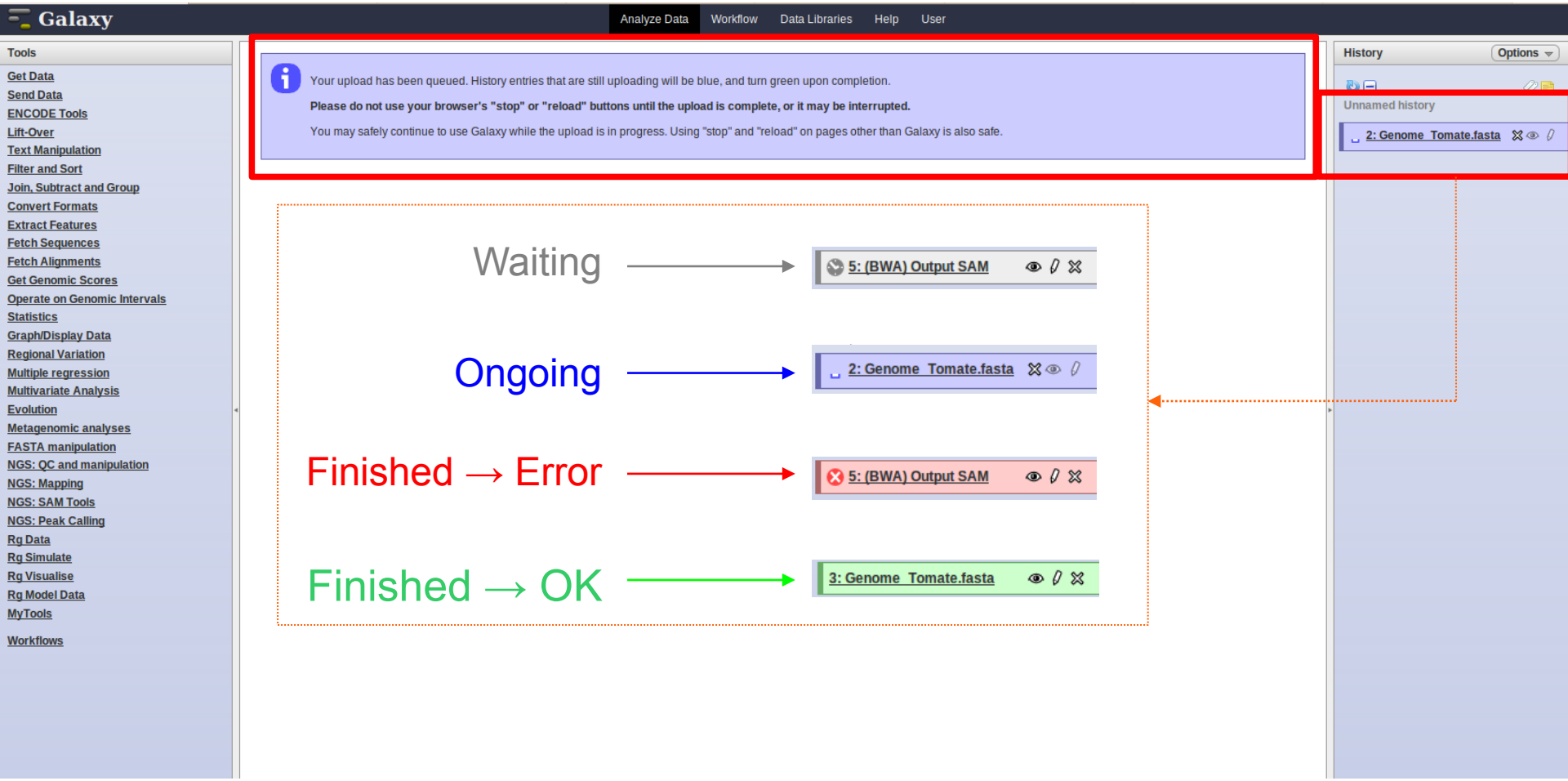
- **FASTX-Toolkit:** tools for FASTA / FASTQ files preprocessing.
- **BWA / Bowtie:** mapping softwares, particularly suitable for short reads alignment (in paired-ends or single-ends) against one reference genome (Burrows – Wheeler Alignement tool).
- **SAMtools:** toolkit for working on the output SAM file (BWA, Bowtie, …).
- **VarScan:** software used to filter SNPs and small indels by:
  - coverage
  - number of variant
  - base quality
  - variant allele frequency
  - pValue

How to upload your data ?

a. Tools → Get Data → Upload File

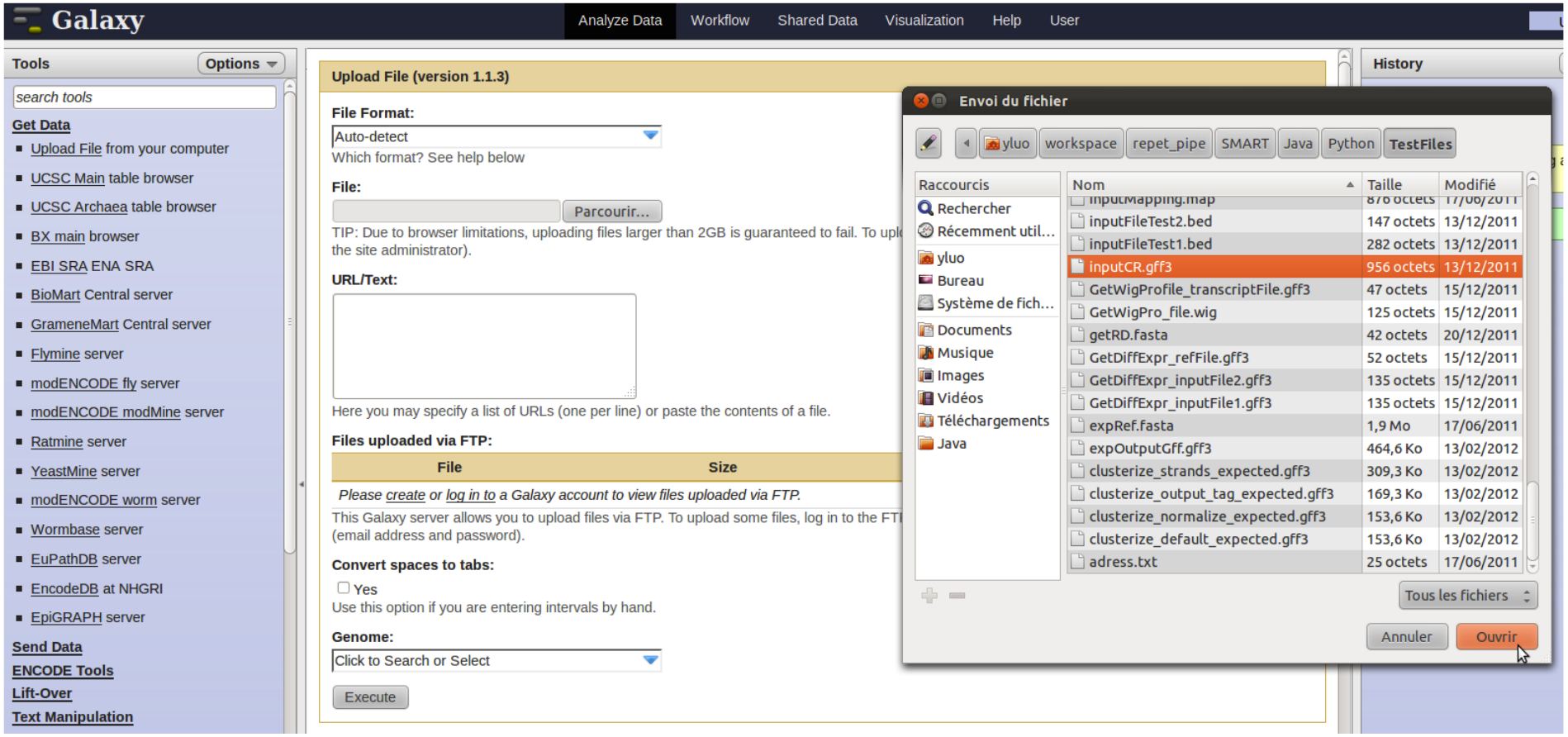


How to upload your data ?

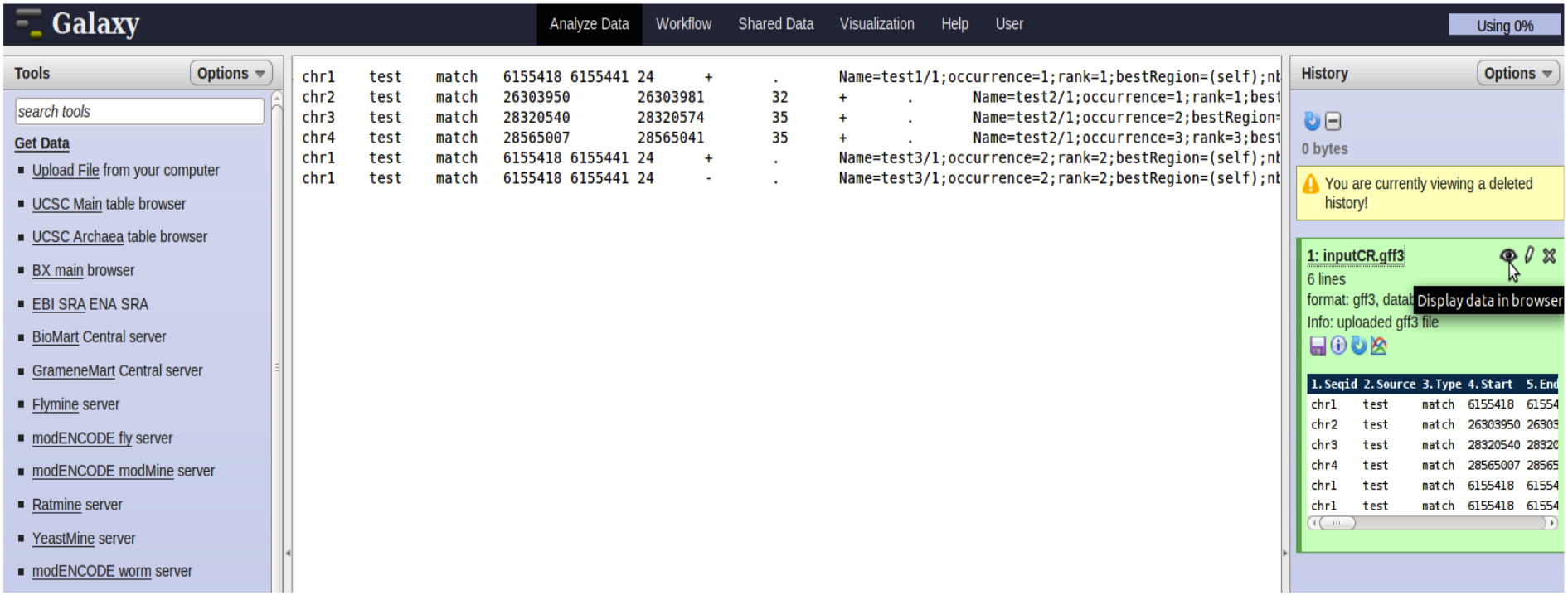


How to upload your data ?

Example

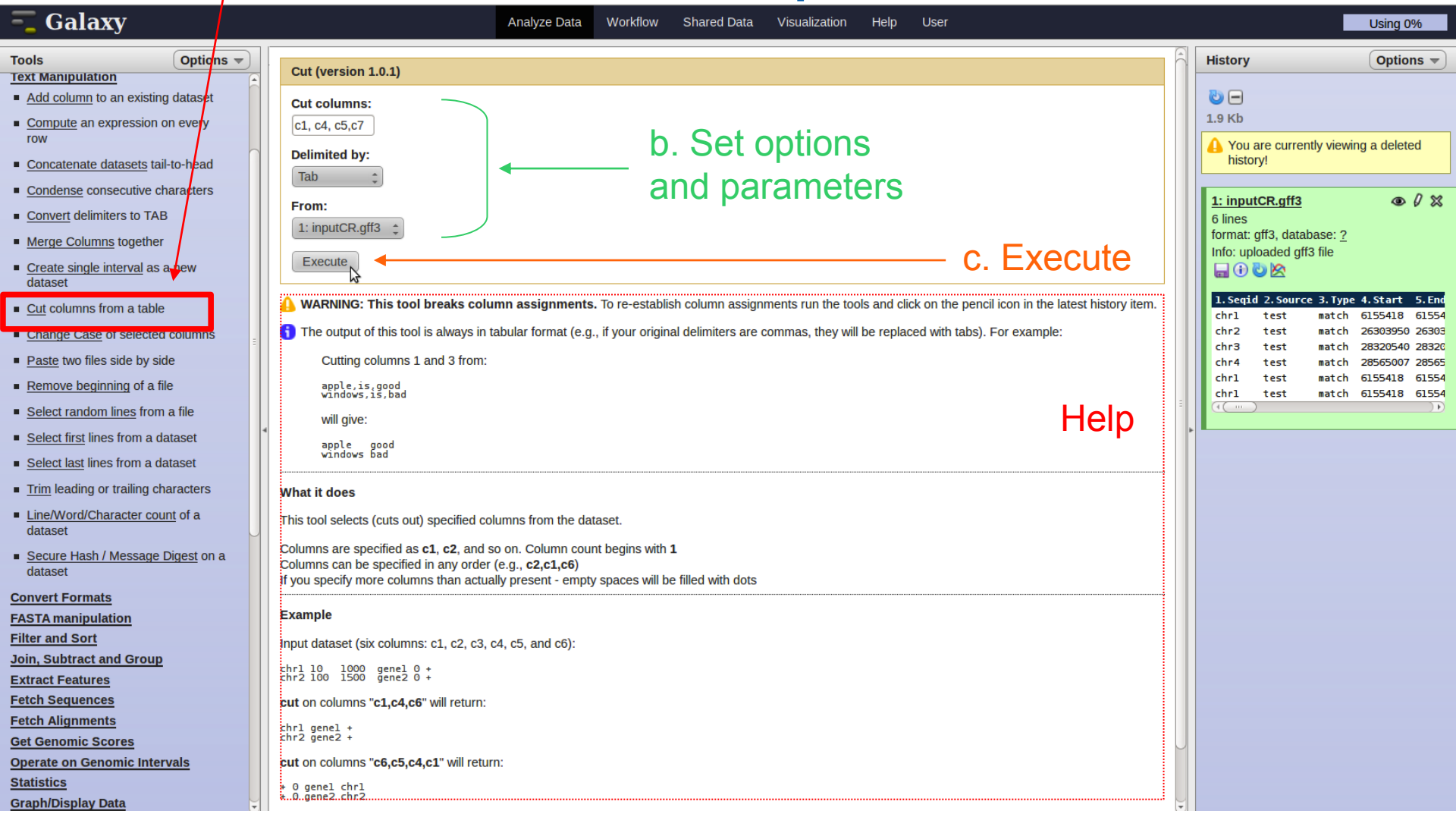


How to upload your data ? Example

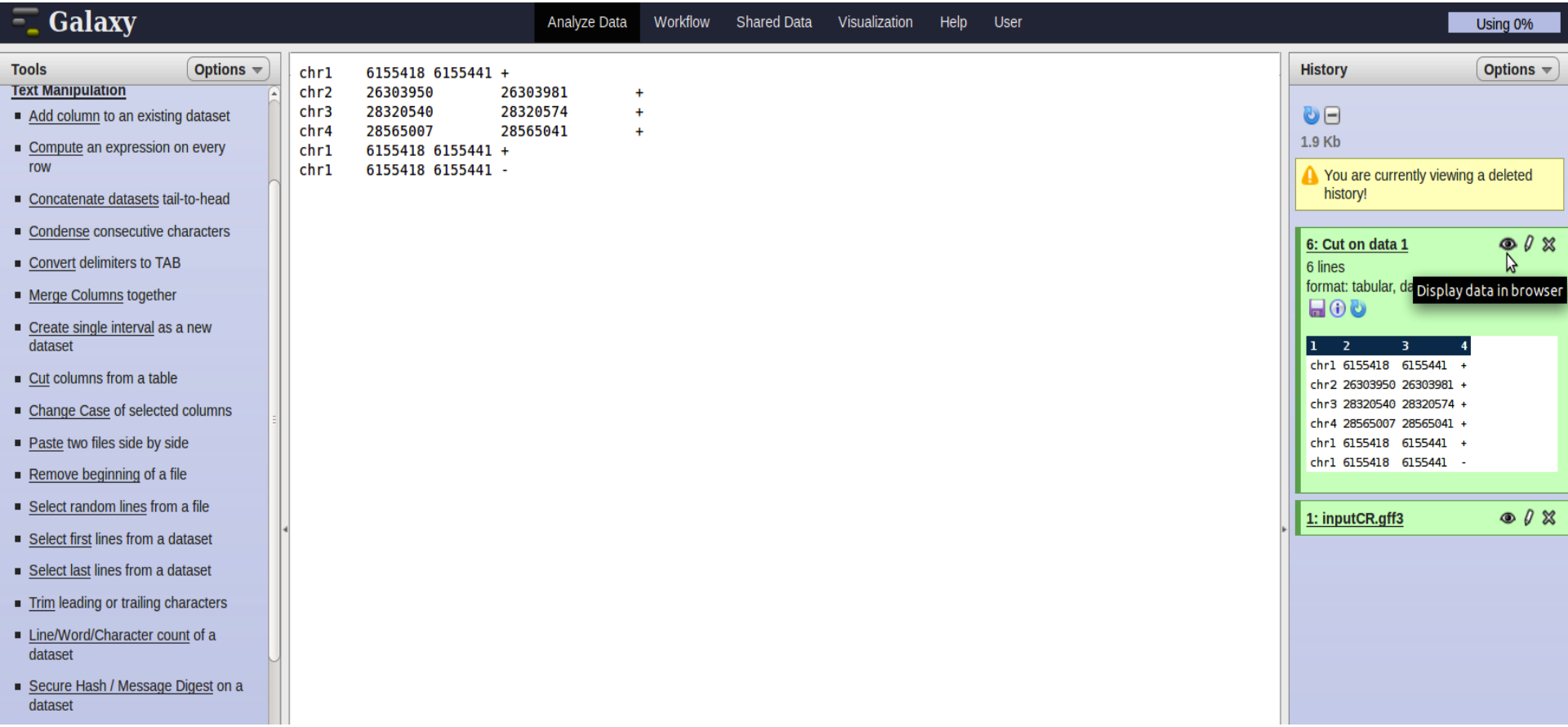


How to use a tool ?

a. Choose a tool from the list **Example**

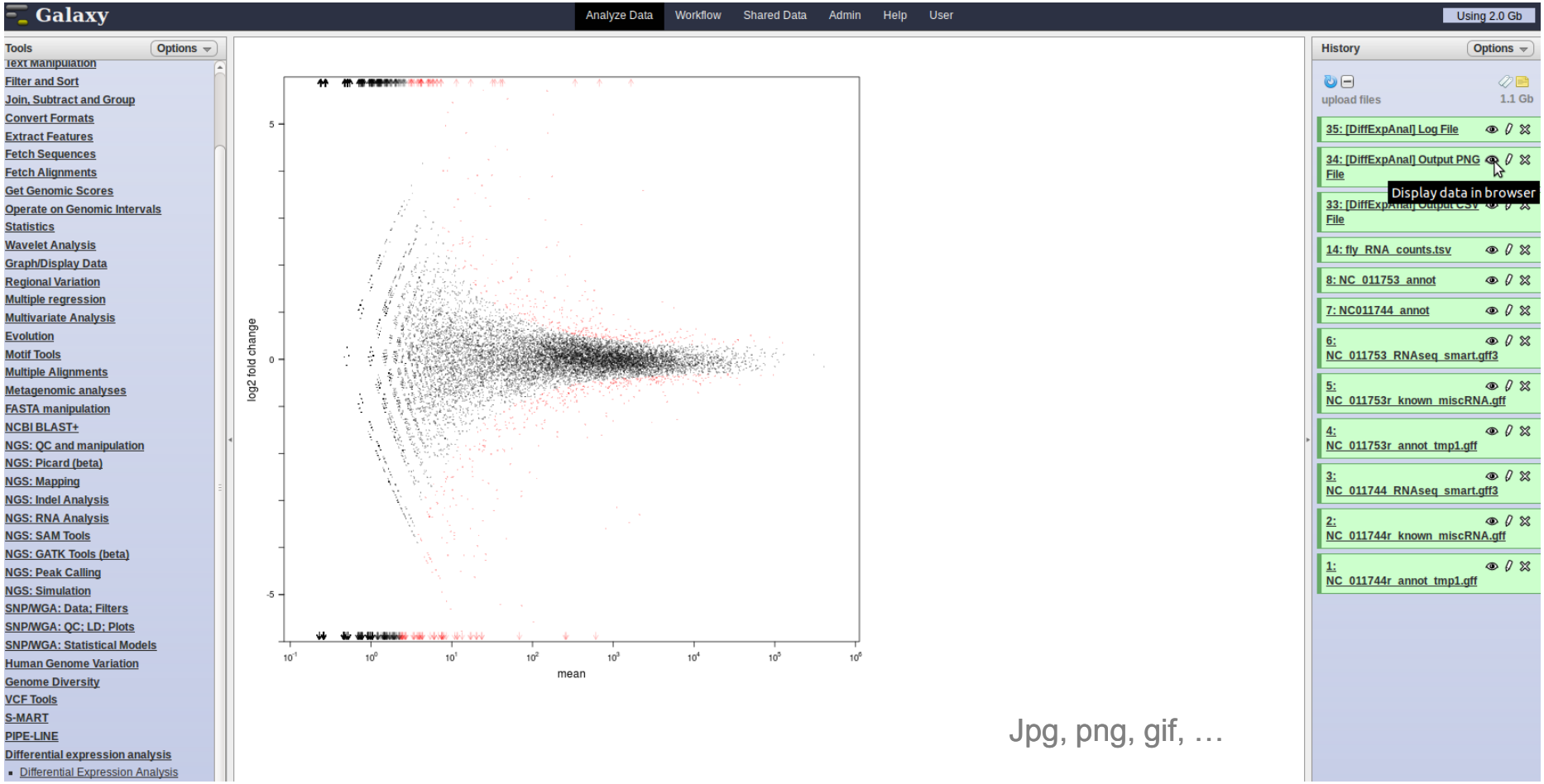


Display results
Example



Display results

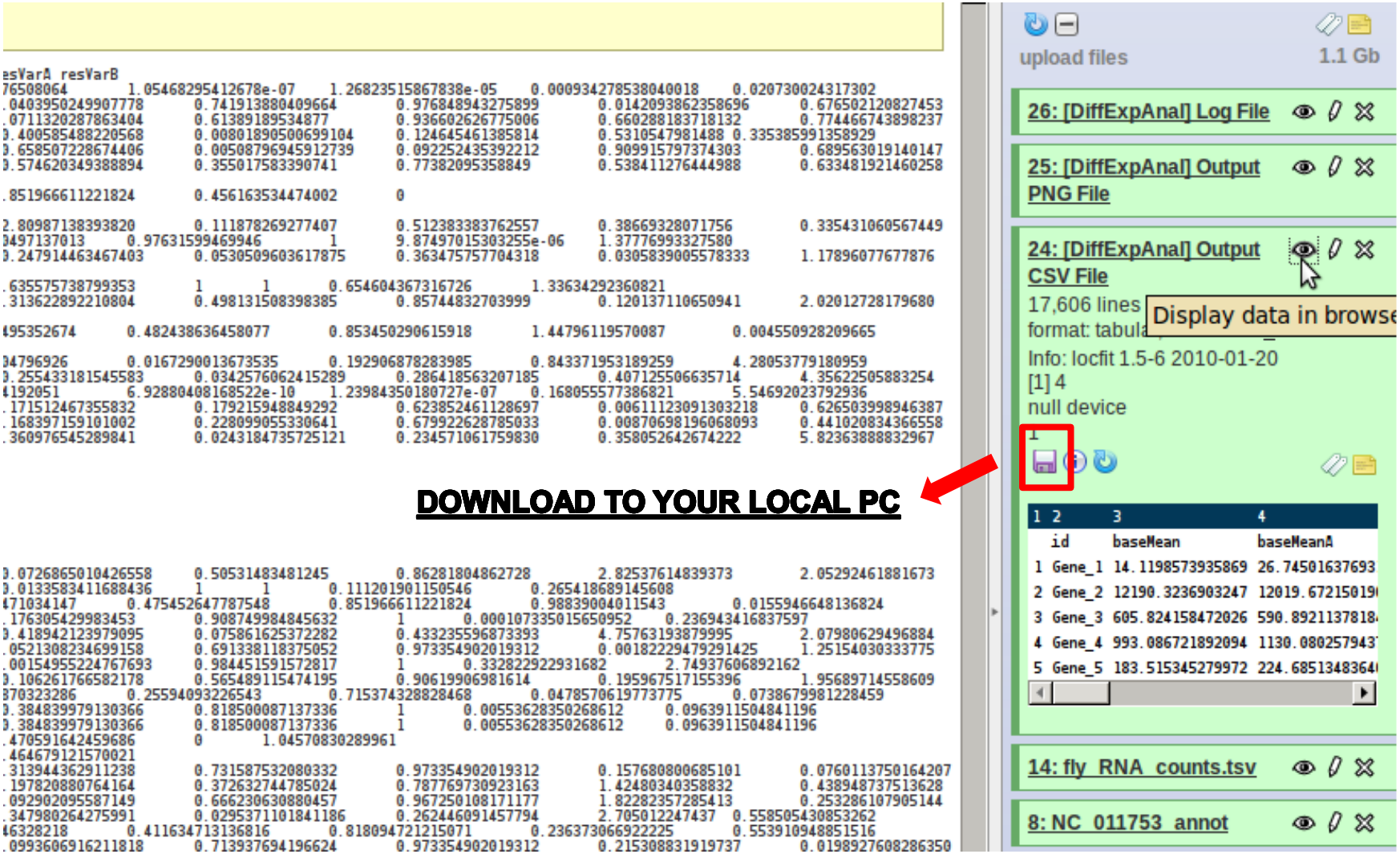
Example



Jpg, png, gif, …

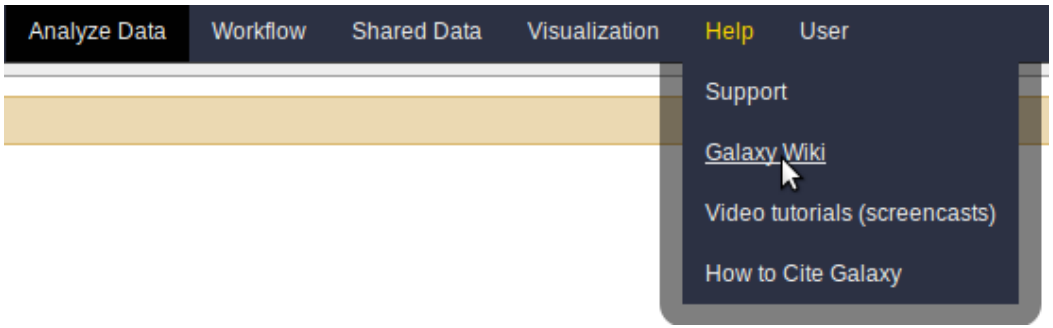
Downdload results

Example



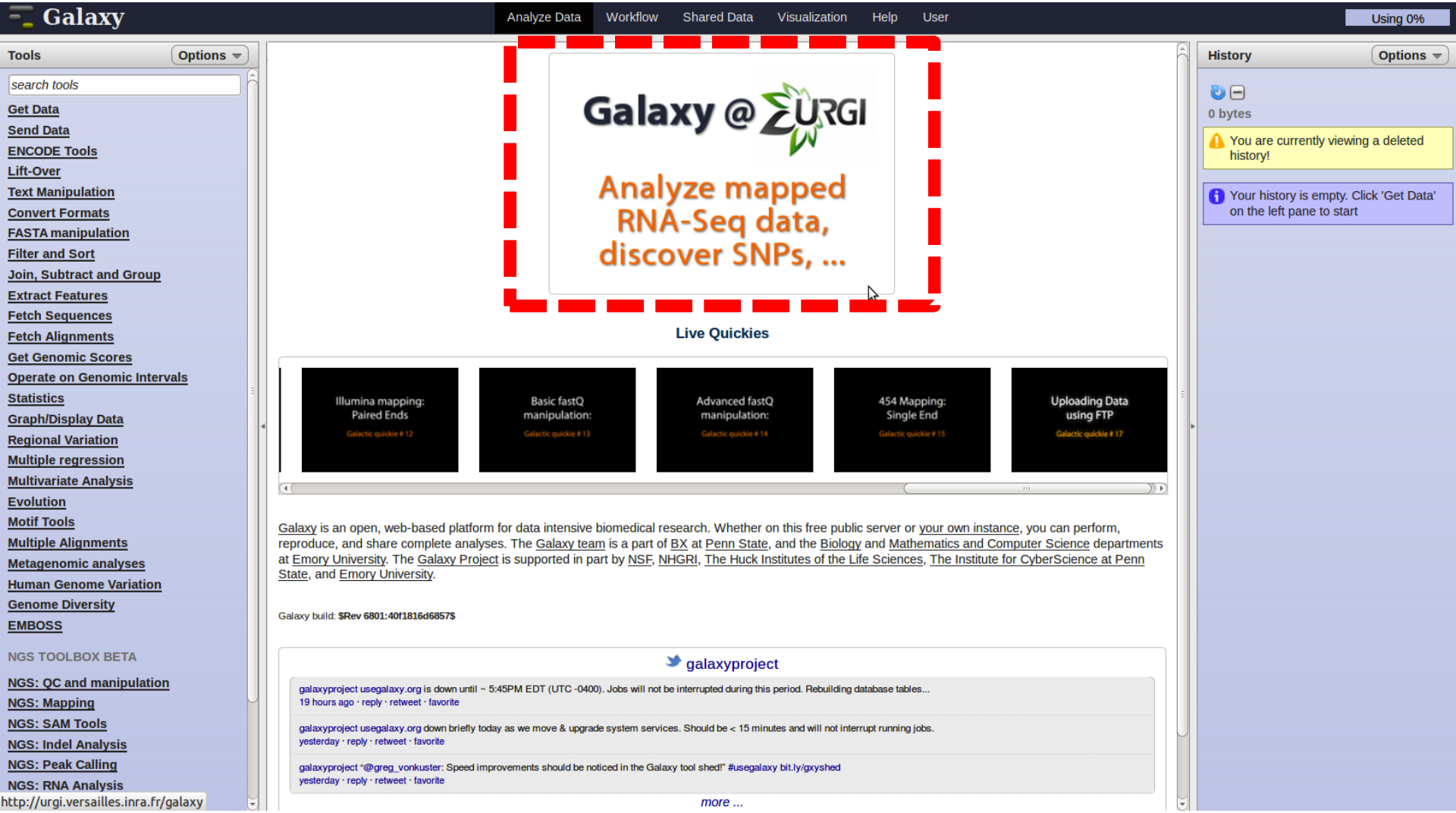
Where can I get help?

- **Galaxy Wiki:**
  http://wiki.g2.bx.psu.edu/



- **Galaxy mailing lists:**
  http://wiki.g2.bx.psu.edu/Mailing%20Lists

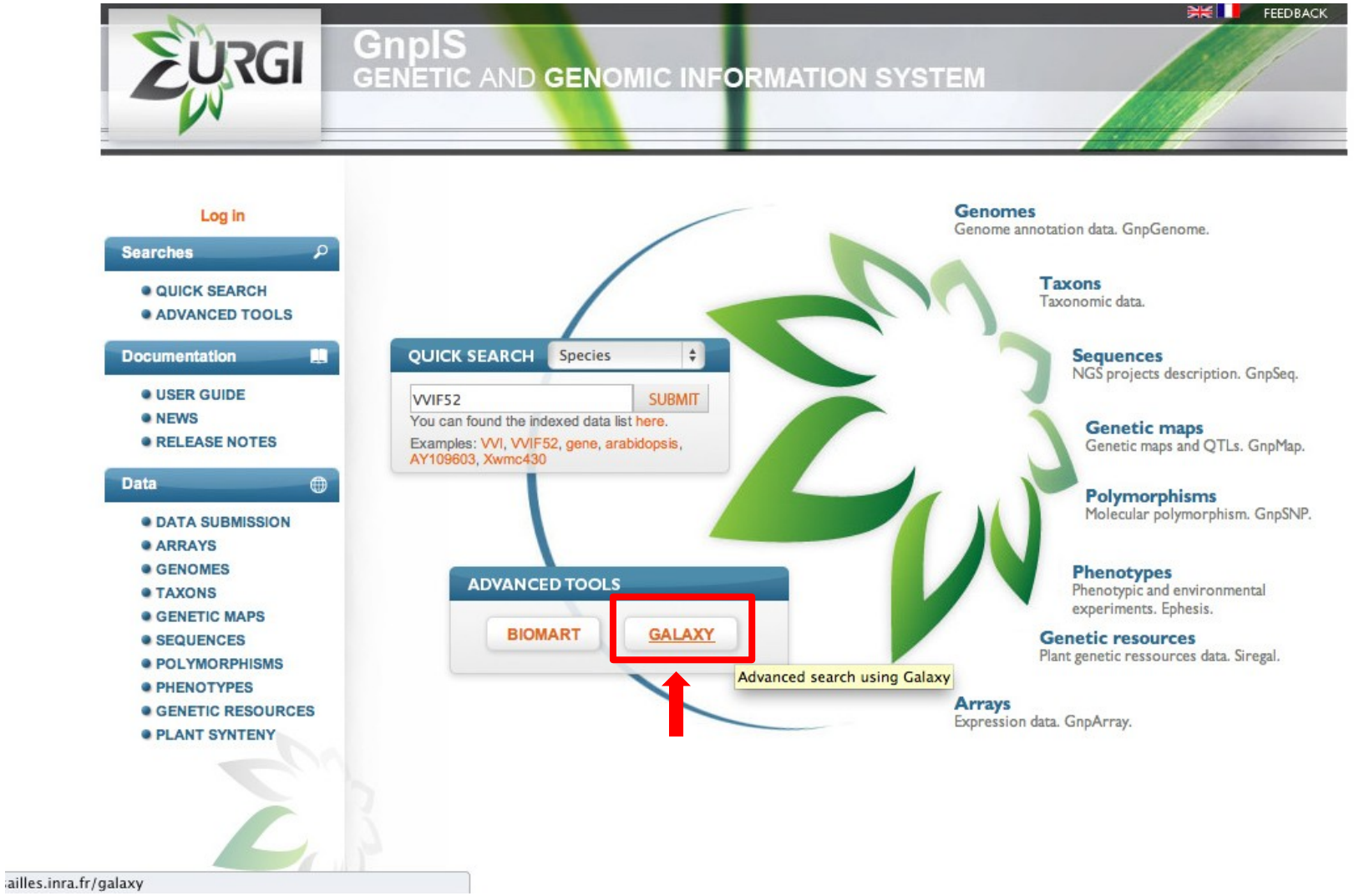
URGI Galaxy


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 0%
Tools
Options
search tools
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
Human Genome Variation
Genome Diversity
EMBOSS
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: Indel Analysis
NGS: Peak Calling
NGS: RNA Analysis
http://urgi.versailles.inra.fr/galaxy
Galaxy @ URGI
Analyze mapped RNA-Seq data, discover SNPs, ...
Live Quickies
Illumina mapping: Paired Ends
Basic fastQ manipulation:
Advanced fastQ manipulation:
454 Mapping: Single End
Uploading Data using FTP
Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.
Galaxy build: $Rev 6801:40f1816d6857$
galaxyproject
galaxyproject usegalaxy.org is down until ~ 5:45PM EDT (UTC -0400). Jobs will not be interrupted during this period. Rebuilding database tables...
19 hours ago · reply · retweet · favorite
galaxyproject usegalaxy.org down briefly today as we move & upgrade system services. Should be < 15 minutes and will not interrupt running jobs.
yesterday · reply · retweet · favorite
galaxyproject "@greg_vonkuster: Speed improvements should be noticed in the Galaxy tool shed!" #usegalaxy bit.ly/gxyshed
yesterday · reply · retweet · favorite
more ...
History
Options
0 bytes
You are currently viewing a deleted history!
Your history is empty. Click 'Get Data' on the left pane to start

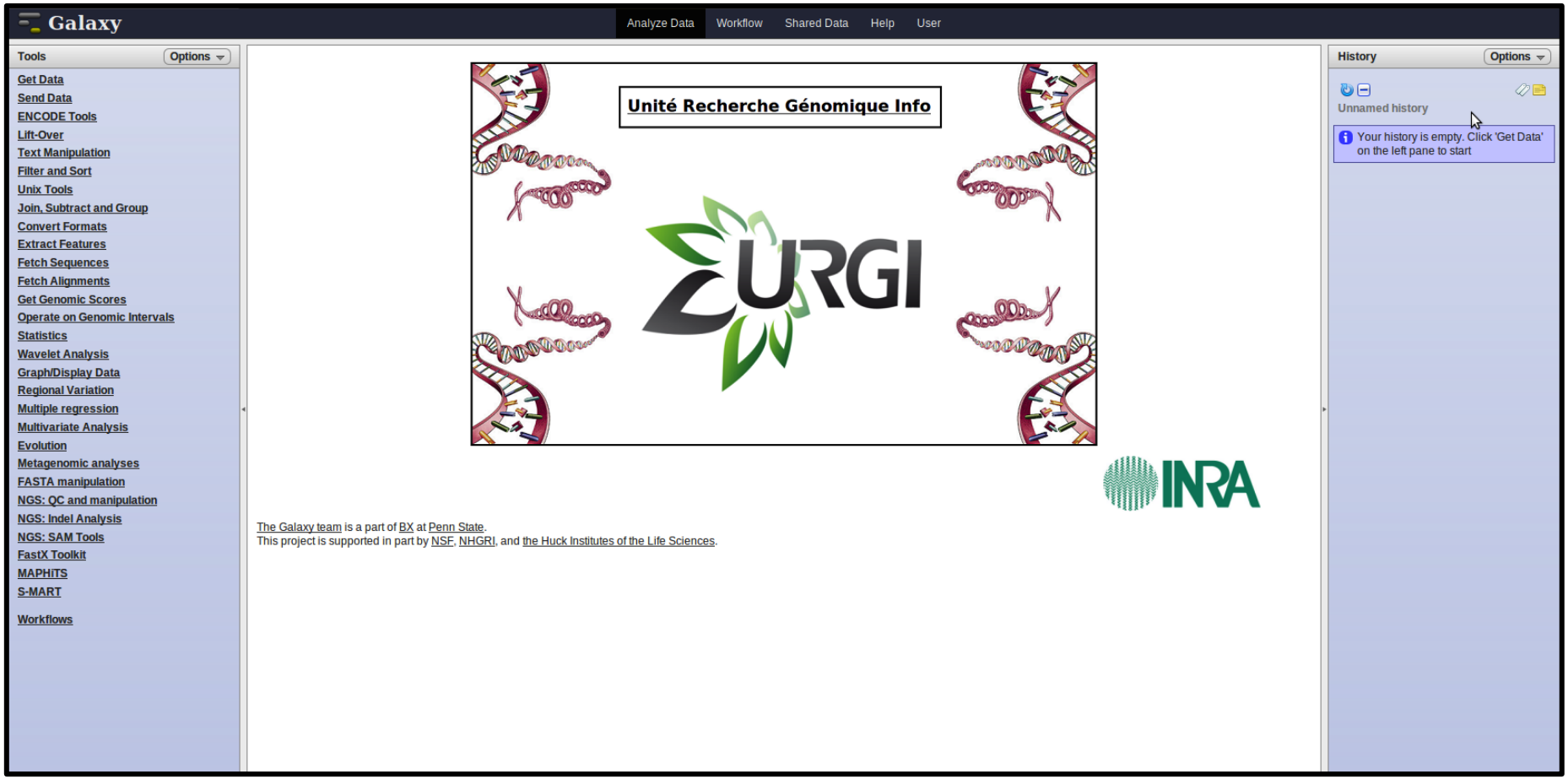
Tools integration in URGI Galaxy

http://urgi.versailles.inra.fr/

Tools integration in URGI Galaxy



Tools integration in URGI Galaxy



Installation of URGI Galaxy

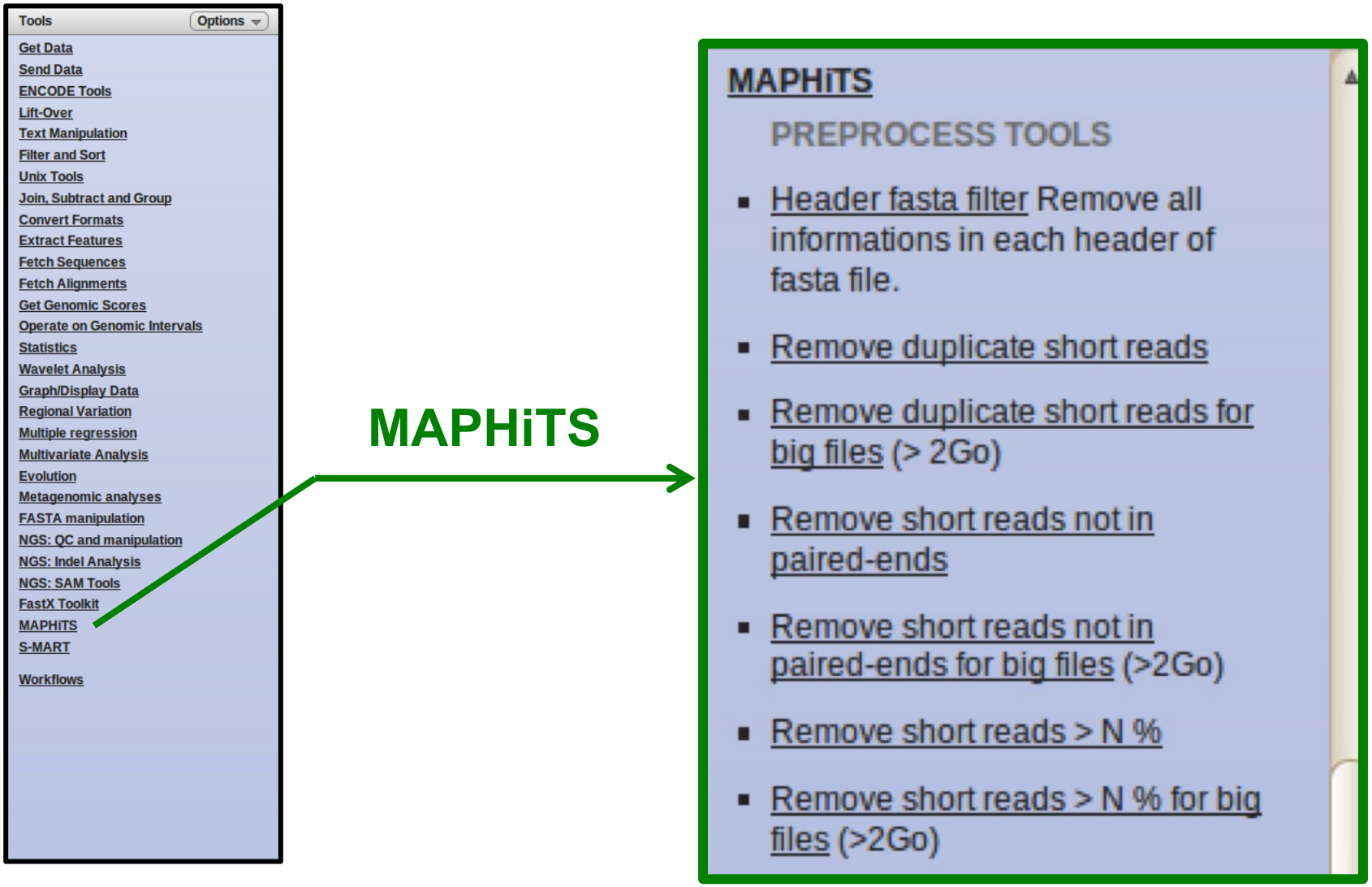
Galaxy is installed on URGI cluster with:

- CPU: **704** (Intel Xeon)
- RAM max: **96 Gb** per job
- Storage: **60 Tb**

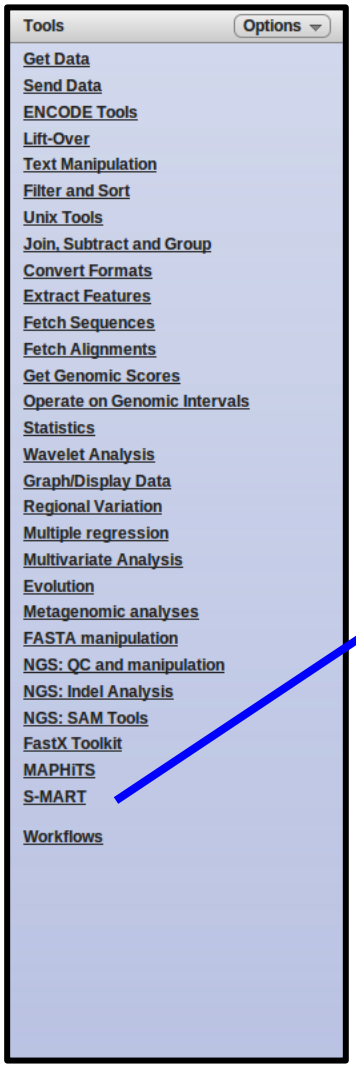
Using Sun Grid Engine (for job managment) and a PostgreSQL Database (for Galaxy).

http://urgi.versailles.inra.fr/galaxy

New URGI Integrated tools



New URGI Integrated tools



S-MART

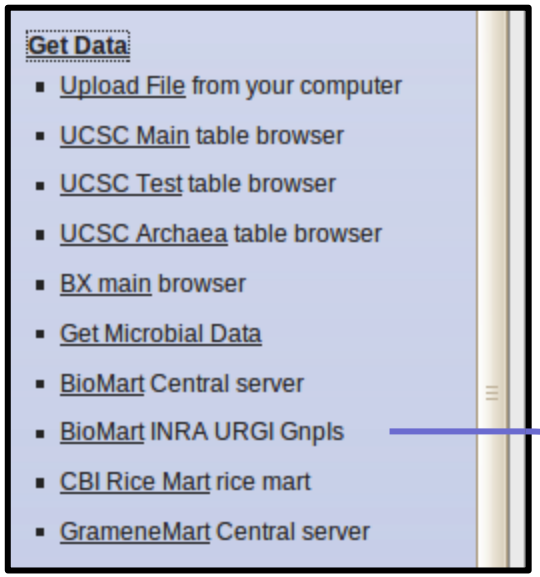
S-MART

CONVERSION TOOLS

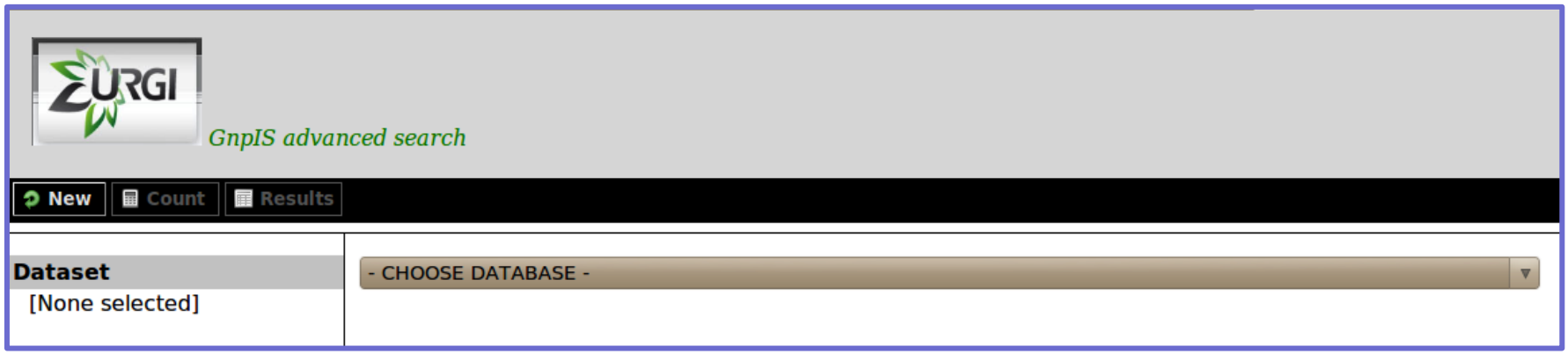
- Gff3 -> Wig Convert Gff3 File to Wig File.
- Bed -> Csv Convert Bed File to Csv File.
- Bed -> Gff2 Convert Bed File to Gff2 File.
- Bed -> Gff3 Convert Bed File to Gff3 File.
- Bed -> Sam Convert Bed File to Sam File.
- Blast (-m 8) -> Csv Convert Blast (-m 8) File to Csv File.
- Blast (-m 8) -> Gff2 Convert Blast (-m 8) File to Gff2 File.
- Blast (-m 8) -> Gff3 Convert Blast (-m 8) File to Gff3 File.

New URGI Integrated tools

Access to URGI Information System via **BioMart** software



BioMart URGI Gnpls



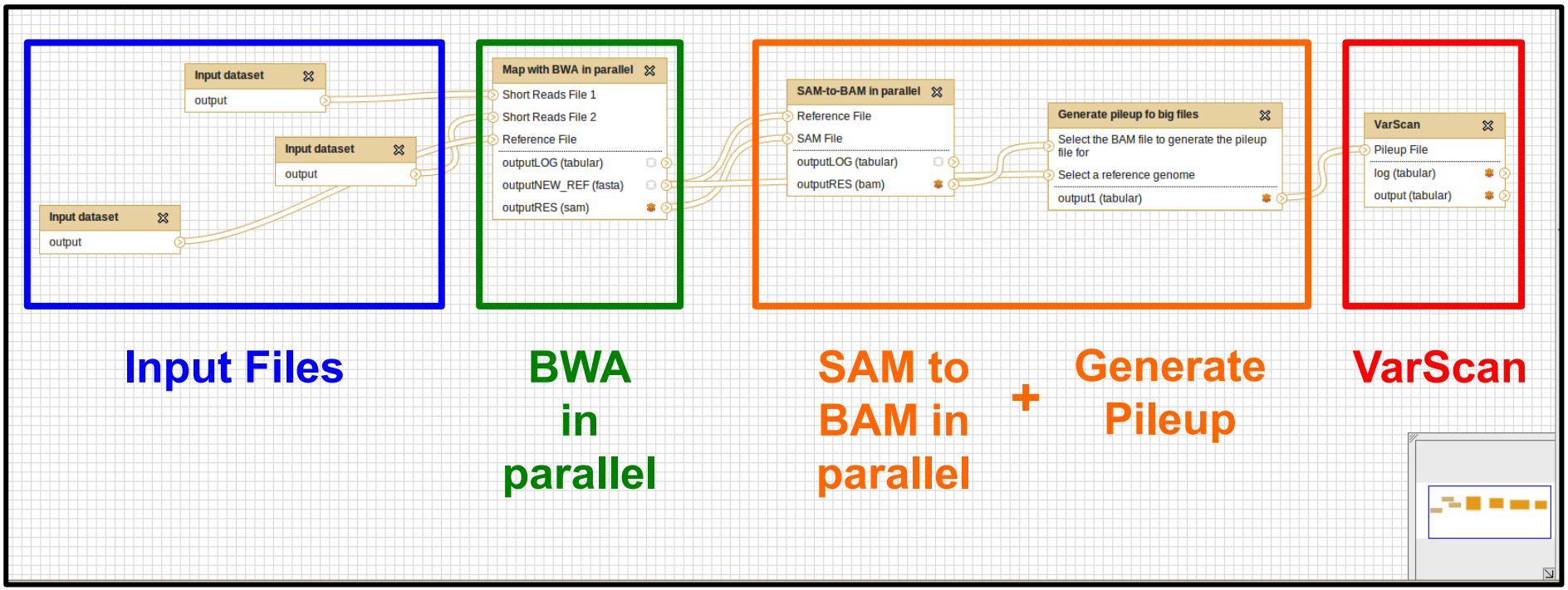
Background and objectives of MAPHiTS

URGI team develops a pipeline *(MAPHiTS)* for SNPs detection from short reads. The set of SNPs is between various species of Grape after mapping short reads against a reference genome.

Link: *http://urgi.versailles.inra.fr/Tools/MAPHiTs*

Contact: *urgi-contact@versailles.inra.fr*

MAPHiTS


Input dataset
output
Input dataset
output
Input dataset
output
Map with BWA in parallel
Short Reads File 1
Short Reads File 2
Reference File
outputLOG (tabular)
outputNEW_REF (fasta)
outputRES (sam)
SAM-to-BAM in parallel
Reference File
SAM File
outputLOG (tabular)
outputRES (bam)
Generate pileup fo big files
Select the BAM file to generate the pileup file for
Select a reference genome
output1 (tabular)
VarScan
Pileup File
log (tabular)
output (tabular)
Input Files
BWA in parallel
SAM to BAM in parallel
+
Generate Pileup
VarScan

Use MAPHiTS in Galaxy

Background and objectives of the S-MART[1]

- RNA-seq (whole transcriptome shotgun sequencing), efficient ways to measure transcriptome data experimentally.

- URGI team develops a tool (S-MART) for mapped RNA-seq data analysis. Such as mapping high-throughput sequencing data from a genome.

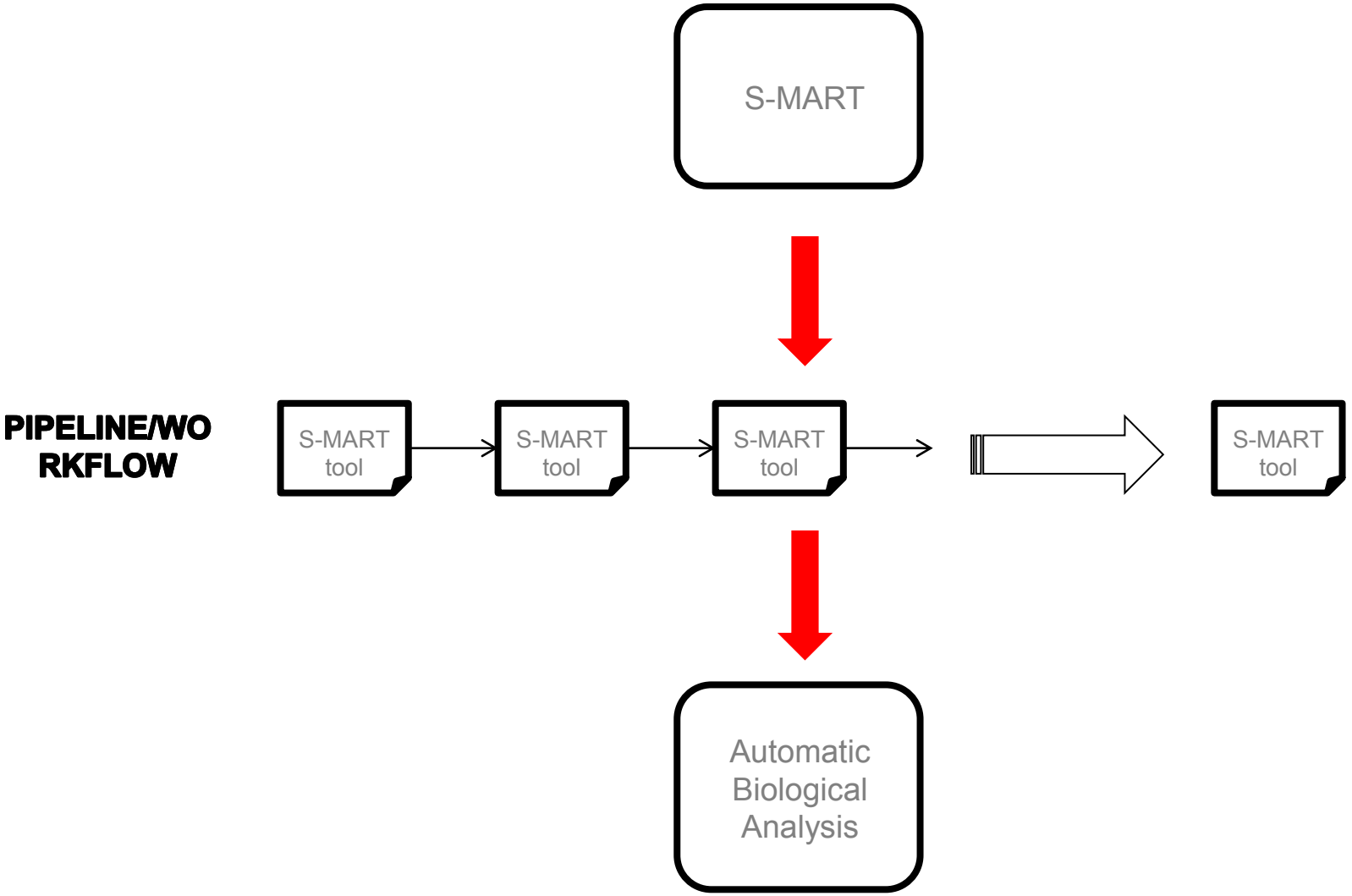
- Use S-MART for data manipulation, visualization, differential expression.

Link: *http://urgi.versailles.inra.fr/Tools/S-MART*

Contact: *matthias.zytnicki@versailles.inra.fr*

1: *Zytnicki M, Quesneville H (2011) S-MART, A Software Toolbox to Aid RNA-seq Data Analaysis. PLoS ONE 6(10):e25988.doi:10.1371/journal.pone.0025988*

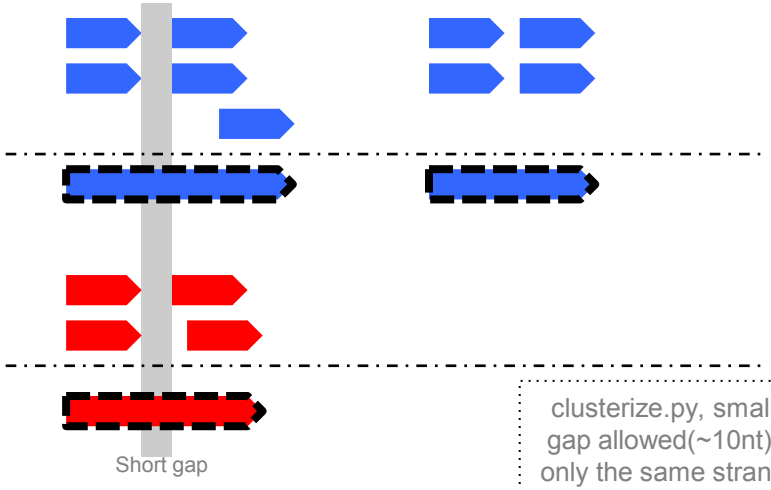
S-MART


S-MART
PIPELINE/WORKFLOW
S-MART tool
S-MART tool
S-MART tool
S-MART tool
Automatic Biological Analysis

S-MART

possible pipeline: differential expression

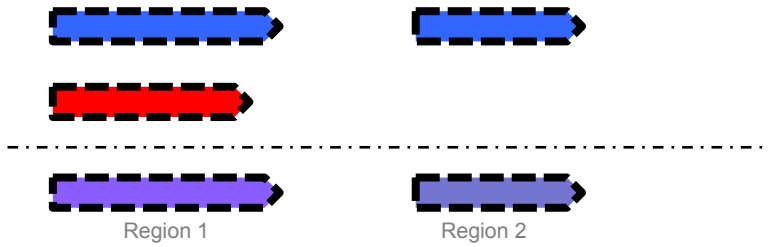
Cluster sample1 and sample2 (different conditions):



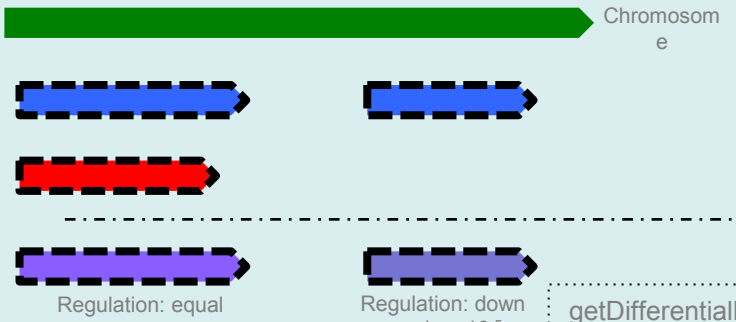
clusterize.py, small gap allowed(~10nt), only the same strand

Merge 2 sets of clusters:

mergeTranscriptLists .py, only the same strand

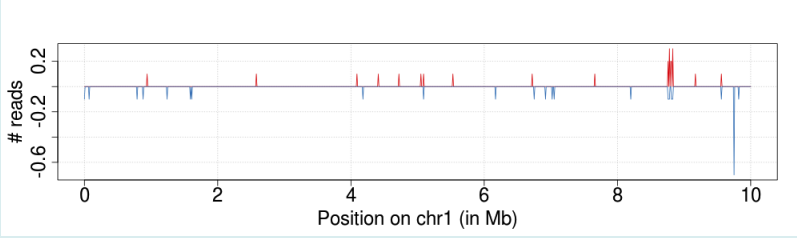


Compute the differential expression:



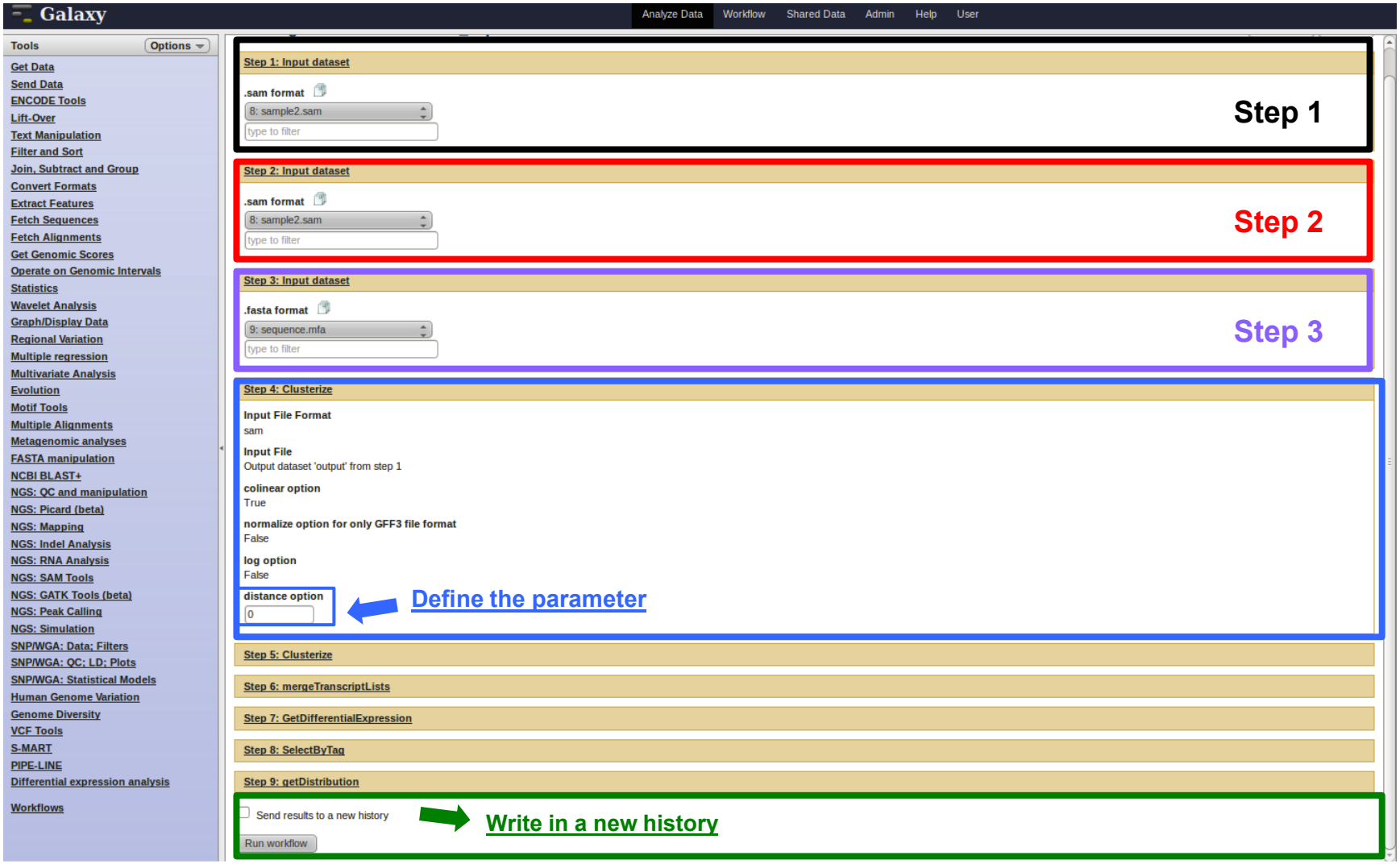
getDifferentialExpression.py, only the same strand

- Select the regions with a very low p-value
- Plot the regions onto the genome.

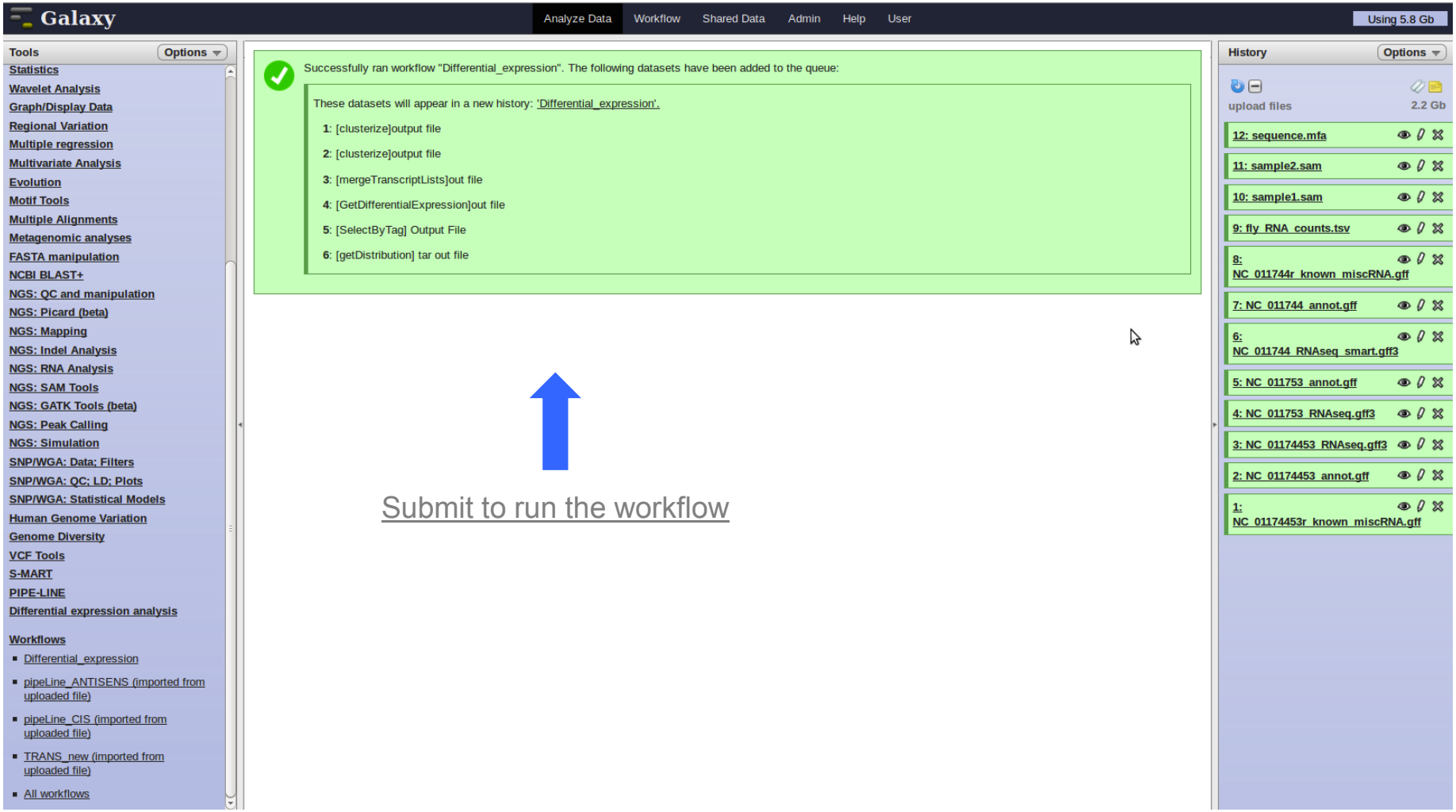


selectByTag.py
getDistribution.py

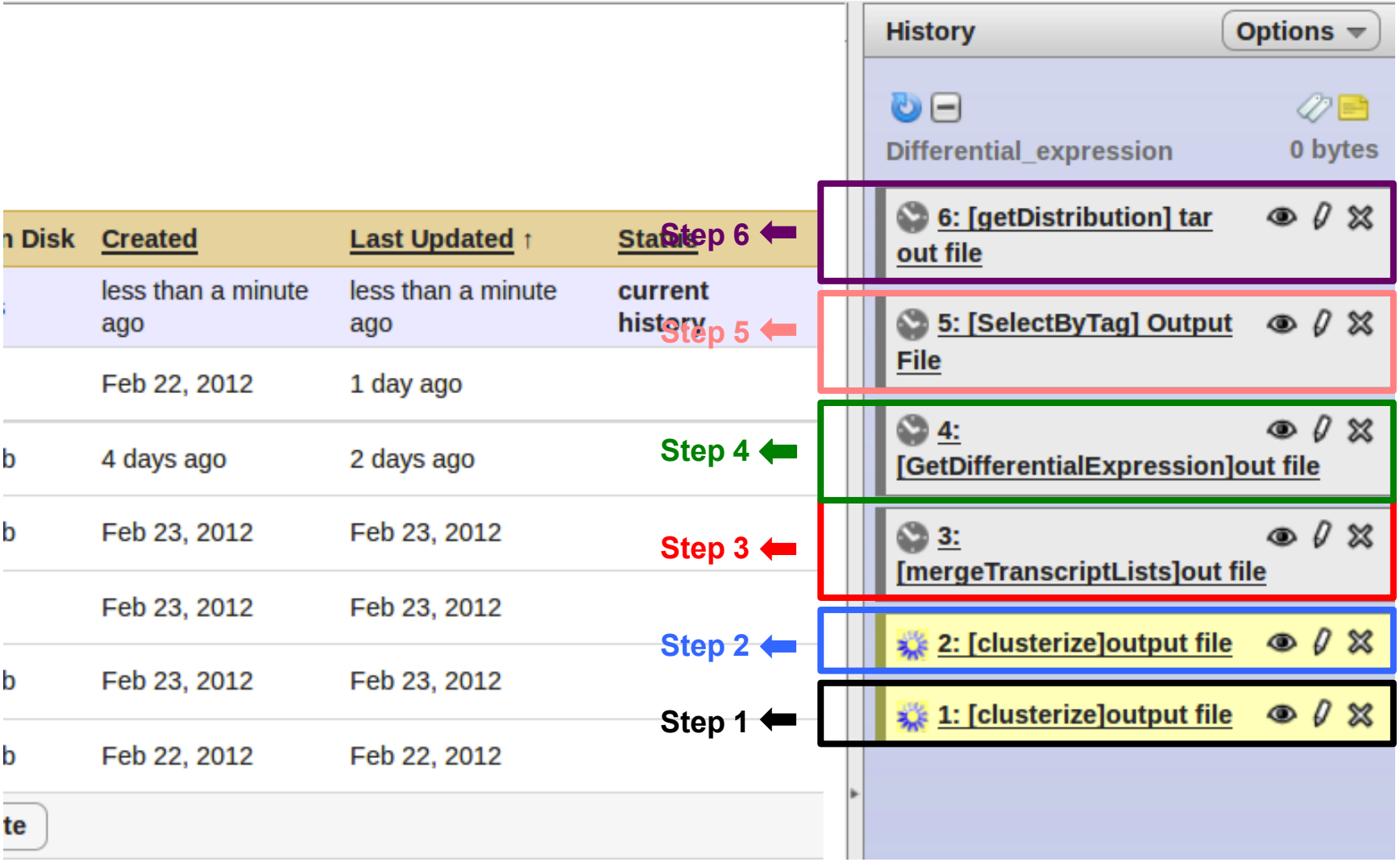
Run S-MART workflow in Galaxy



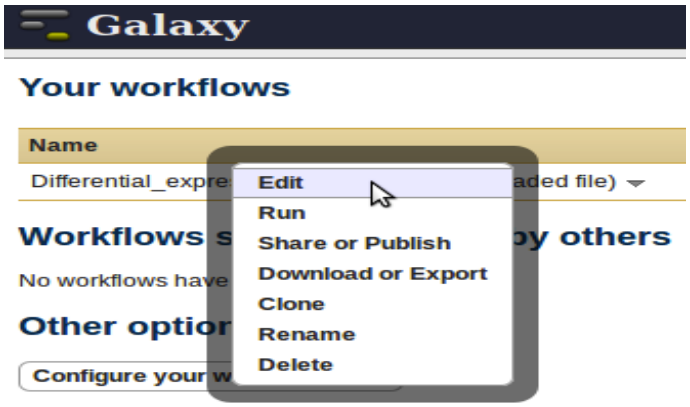
Run S-MART workflow in Galaxy

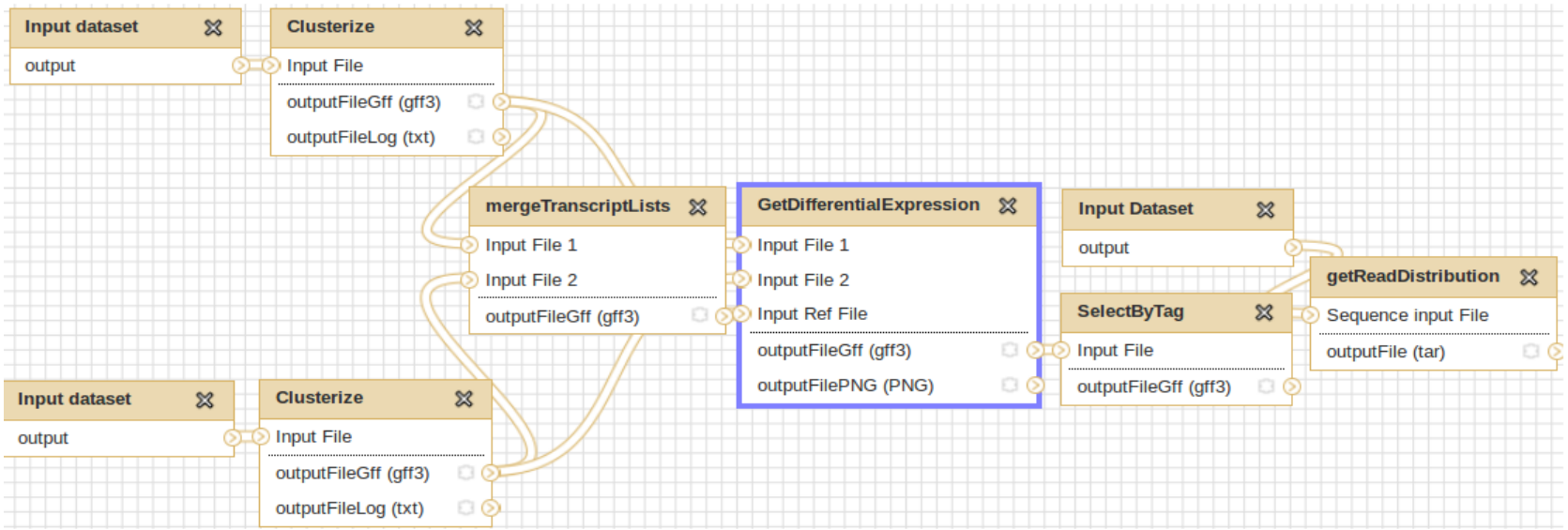


Run S-MART workflow in Galaxy

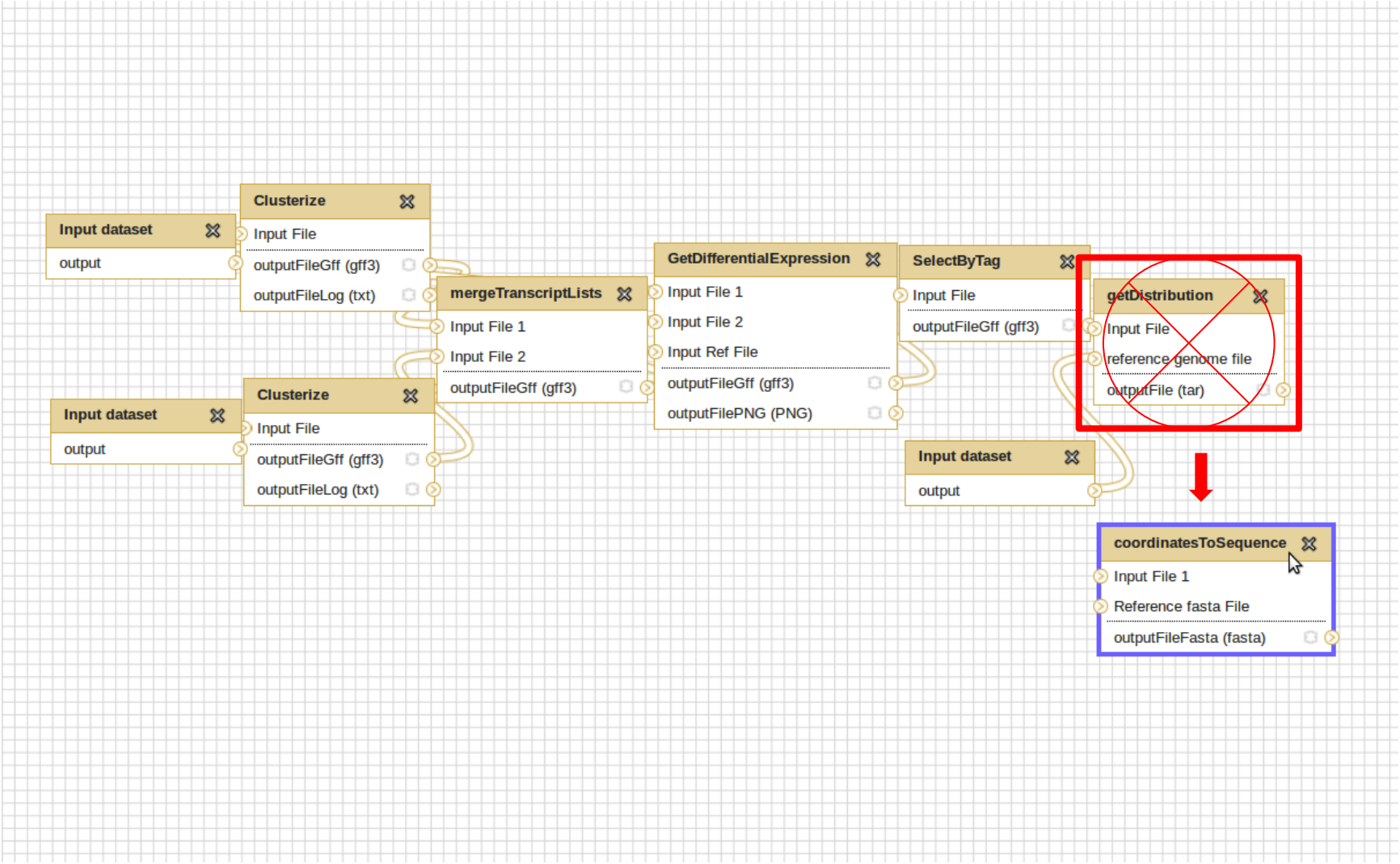


Edit S-MART workflow in Galaxy

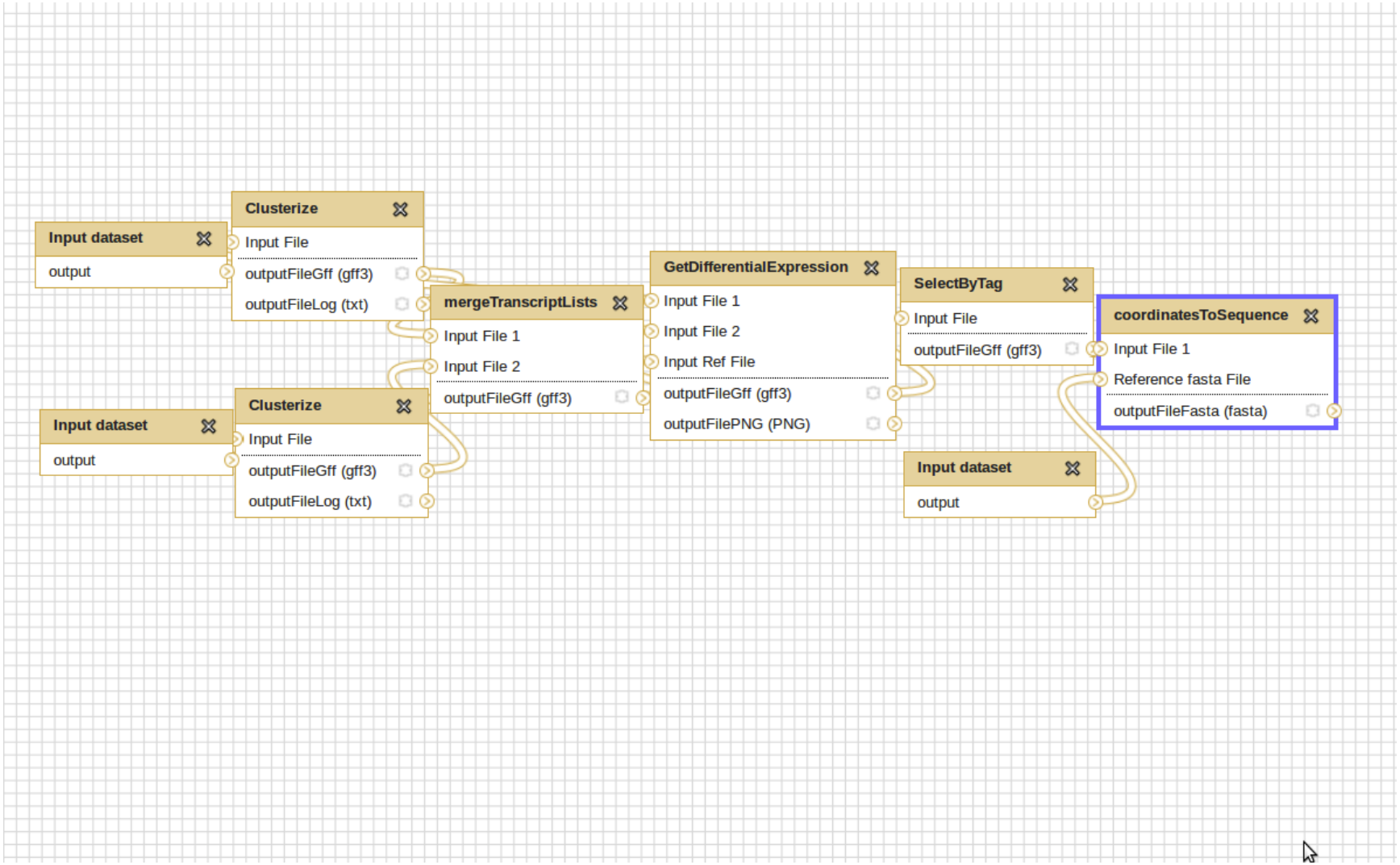




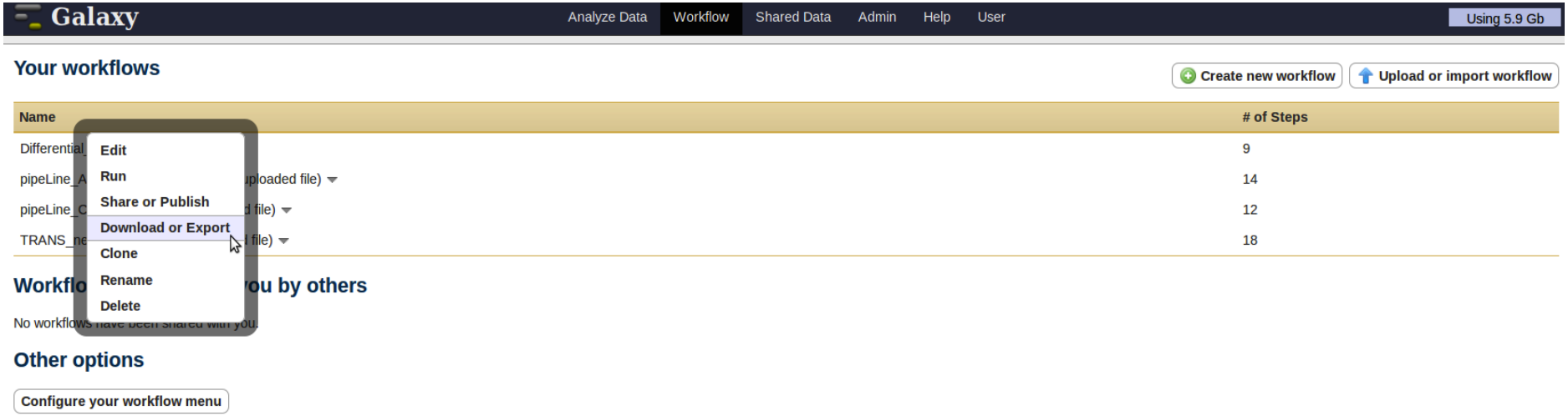
Edit S-MART workflow in Galaxy



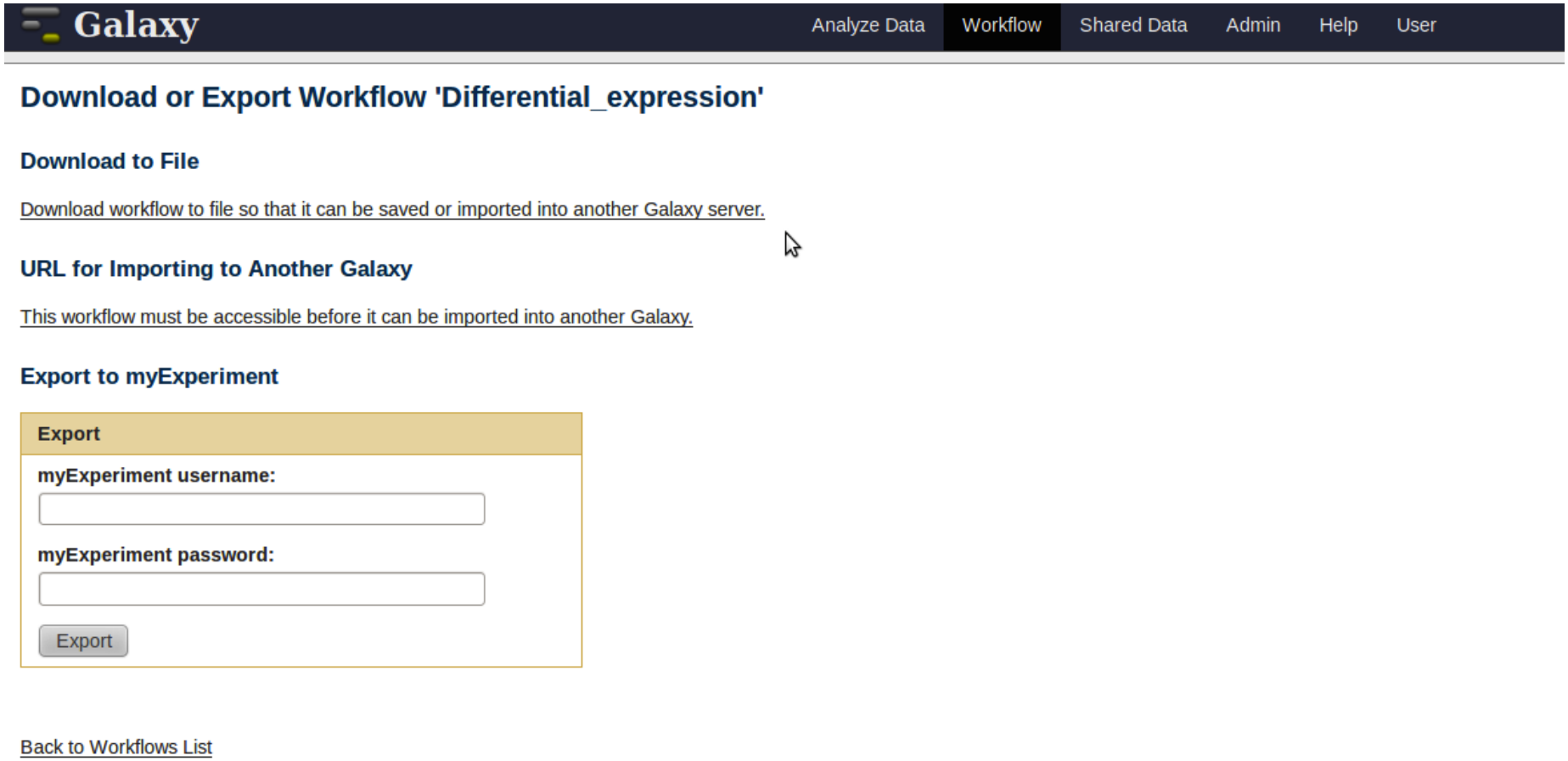
Edit S-MART workflow in Galaxy



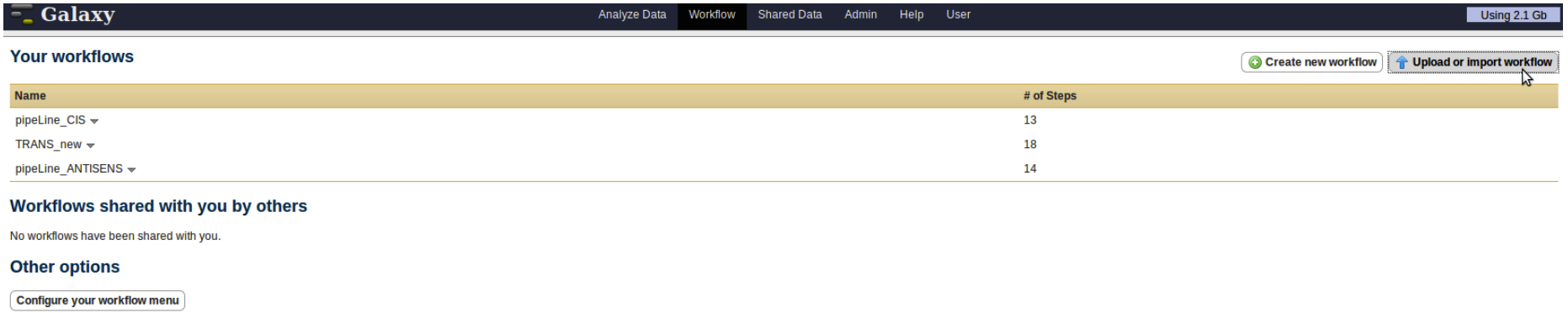
Export S-MART workflow from Galaxy



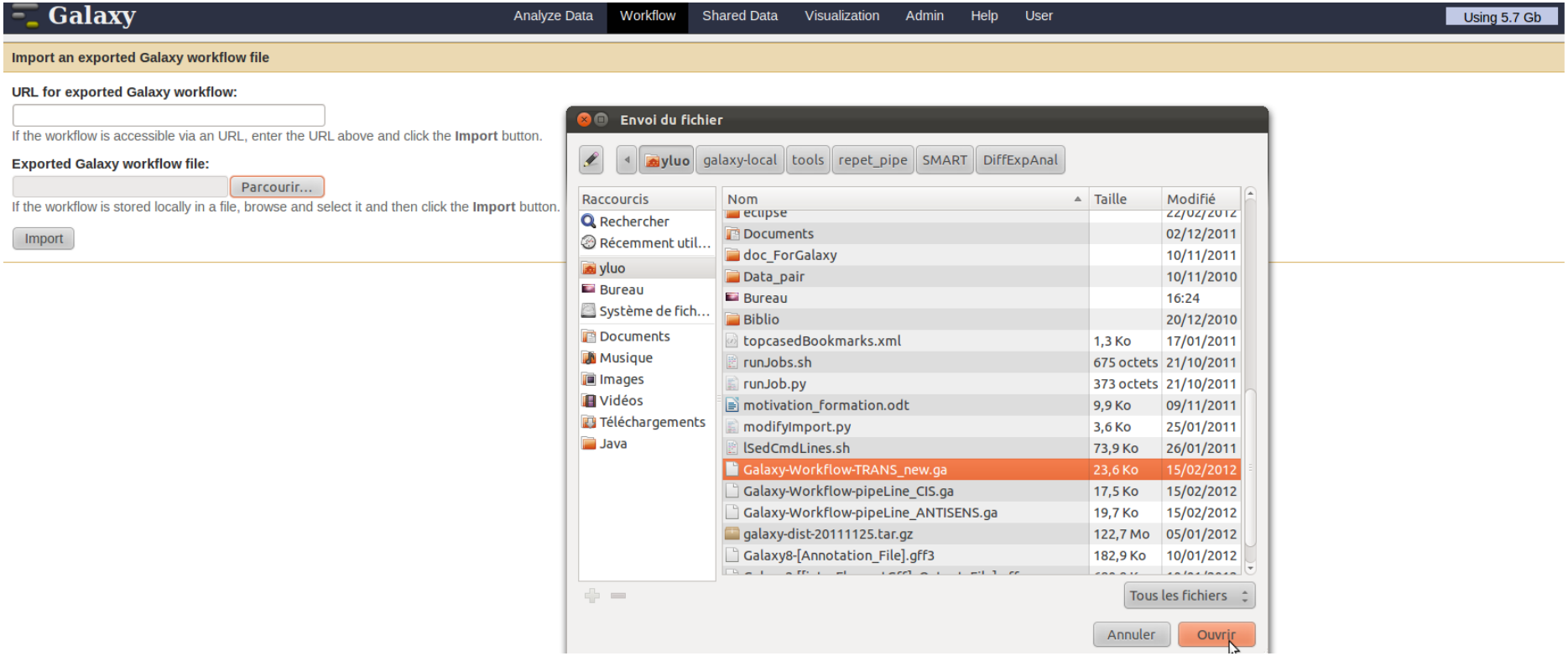
Export S-MART workflow from Galaxy



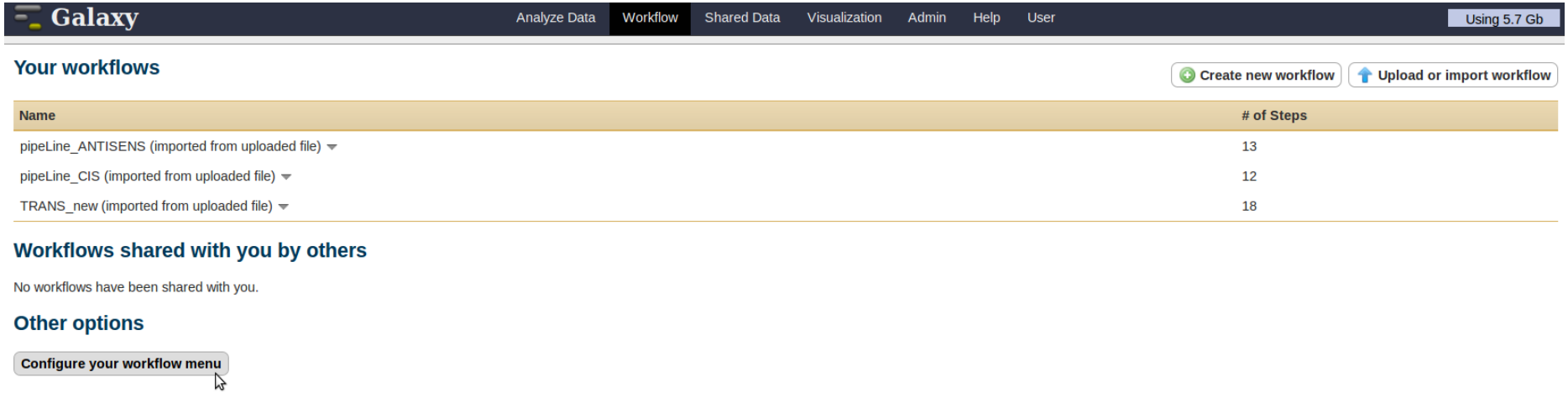
Import S-MART workflow in Galaxy



Import S-MART workflow in Galaxy



Display S-MART workflow in Galaxy



Display S-MART workflow in Galaxy

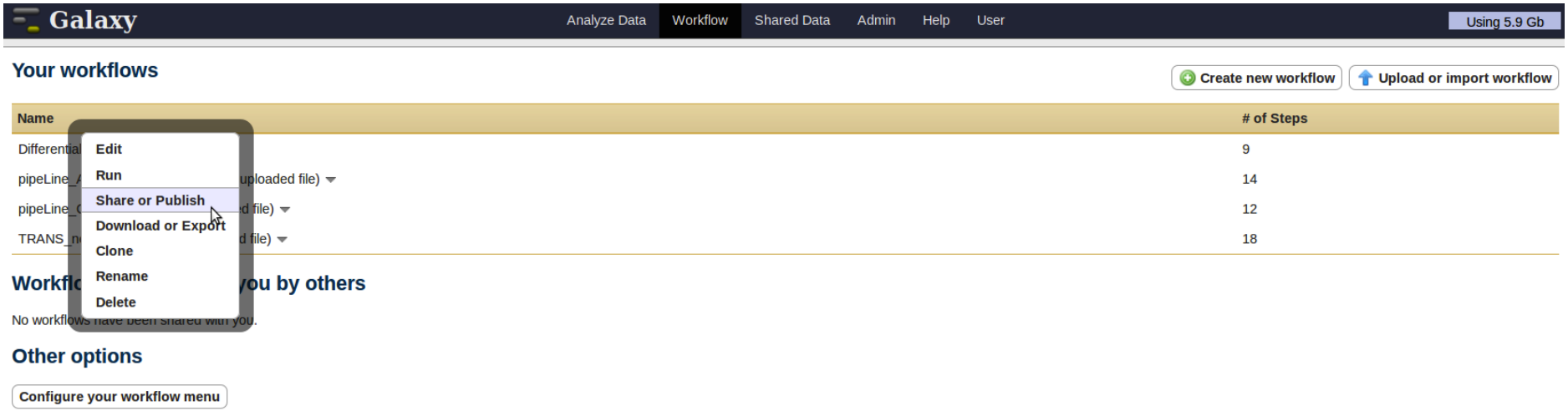
Name	Owner	# of Steps	Show in menu
Differential_expression	You	9	☑
pipeLine_ANTISENS (imported from uploaded file)	You	14	☑
pipeLine_CIS (imported from uploaded file)	You	12	☑
TRANS_new (imported from uploaded file)	You	18	☑

Envoyer

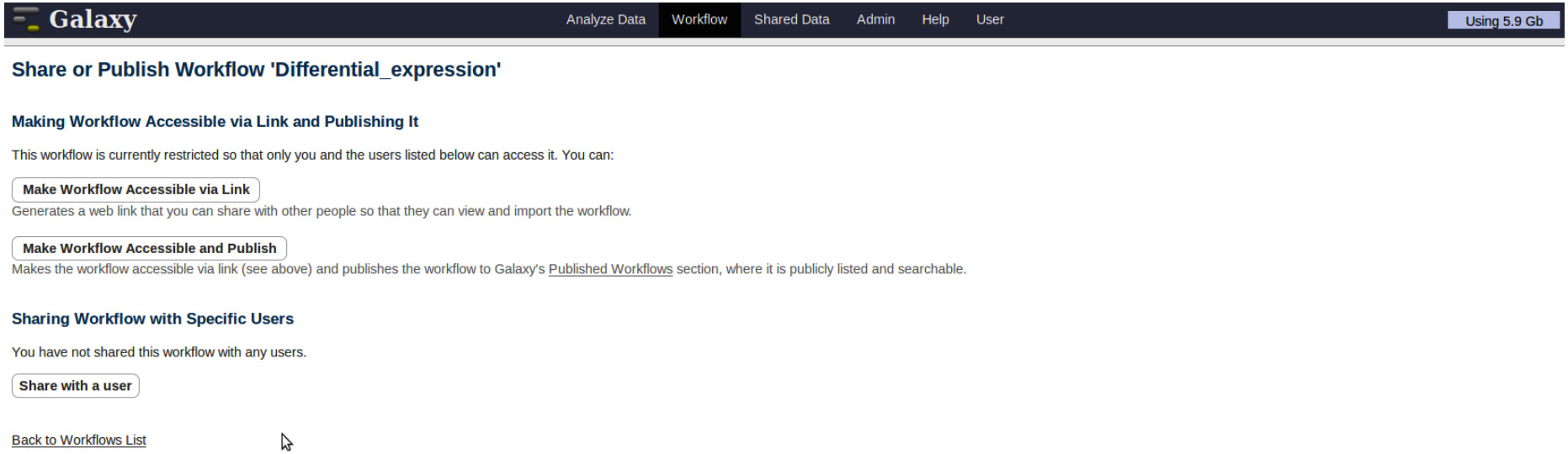
Workflows

- Differential_expression
- pipeLine_ANTISENS (imported from uploaded file)
- pipeLine_CIS (imported from uploaded file)
- TRANS_new (imported from uploaded file)
- All workflows

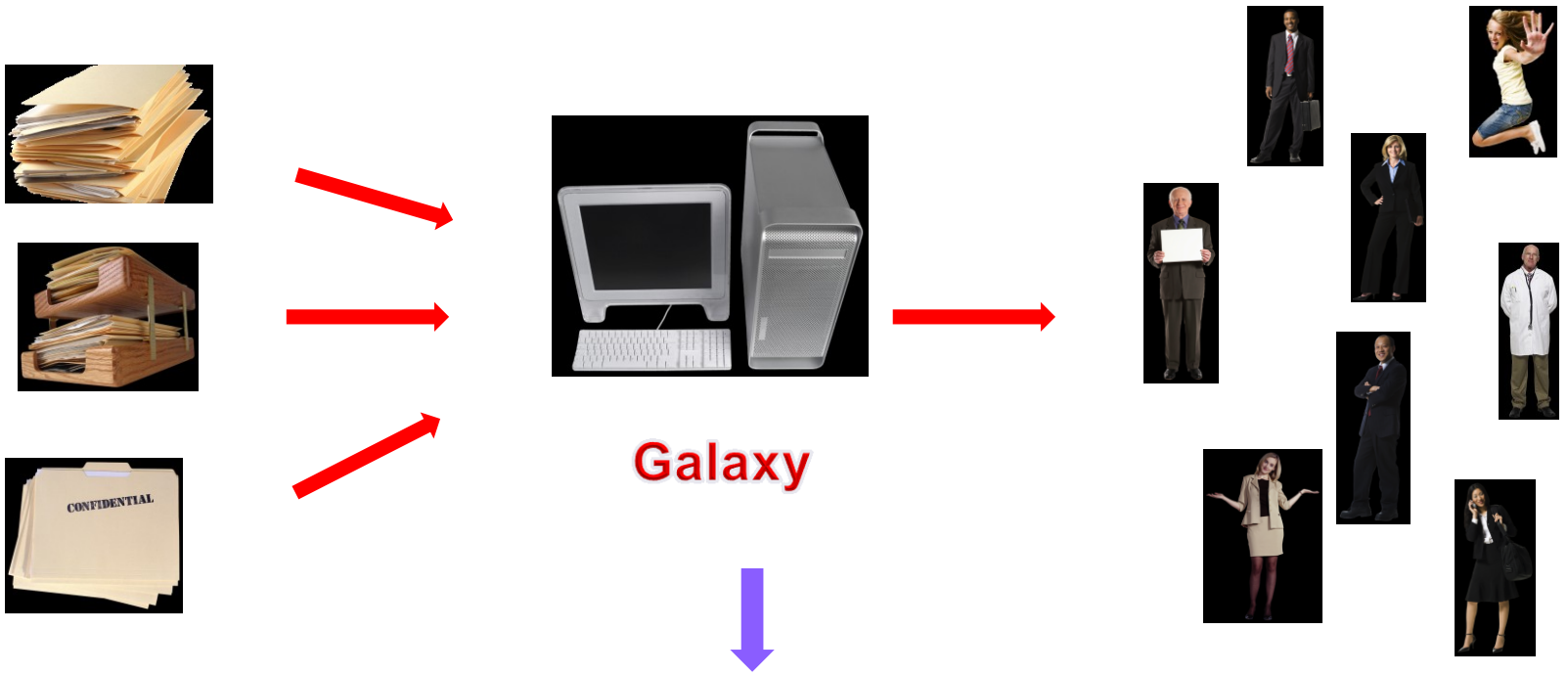
Share S-MART workflow in Galaxy



Share S-MART workflows in Galaxy



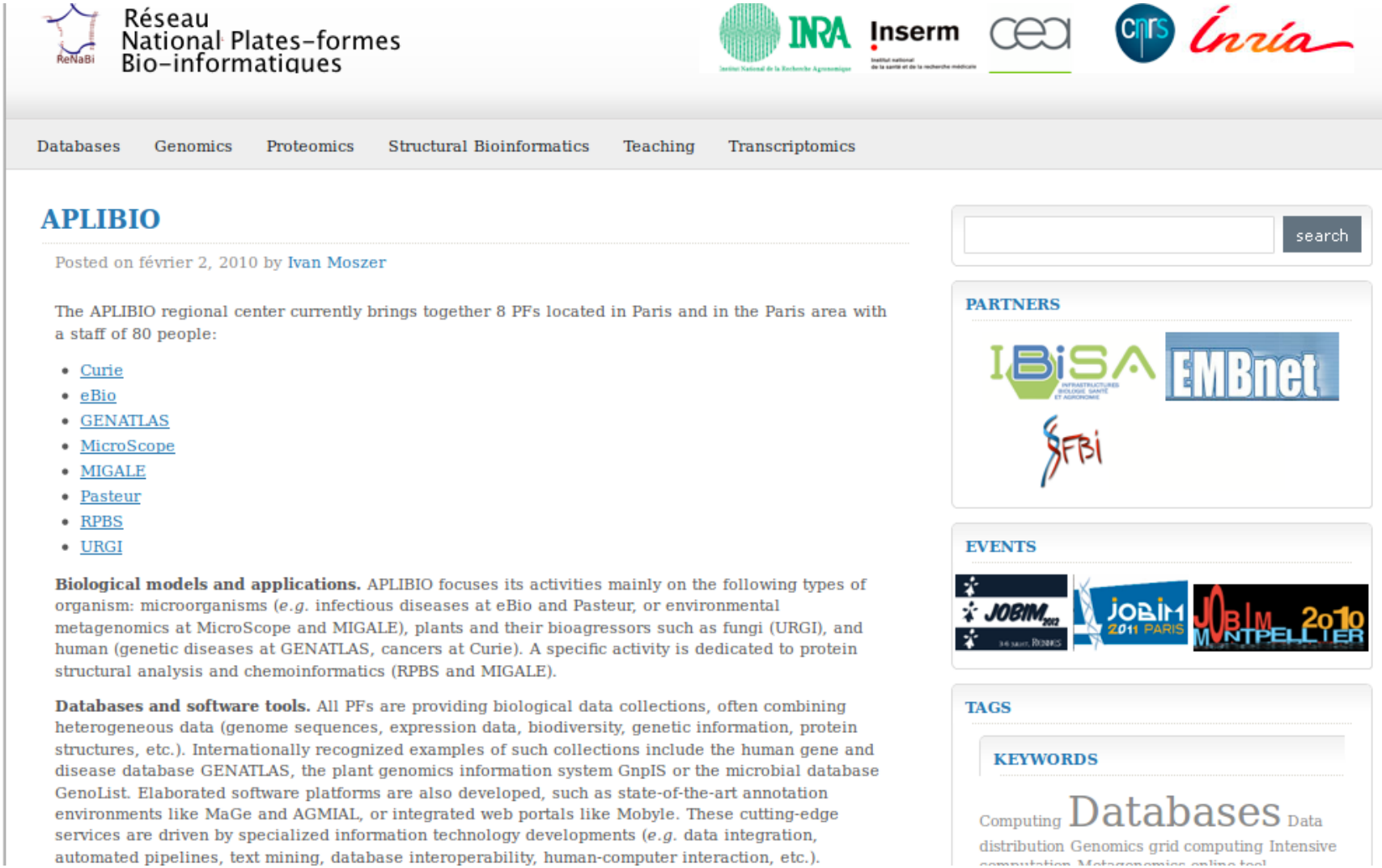
Share workflows & analysis tools in Galaxy



APLIBIO Projects (IBISA funding)

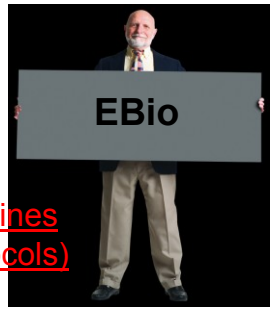
APLIBIO

http://www.renabi.fr/platforms/aplibio/



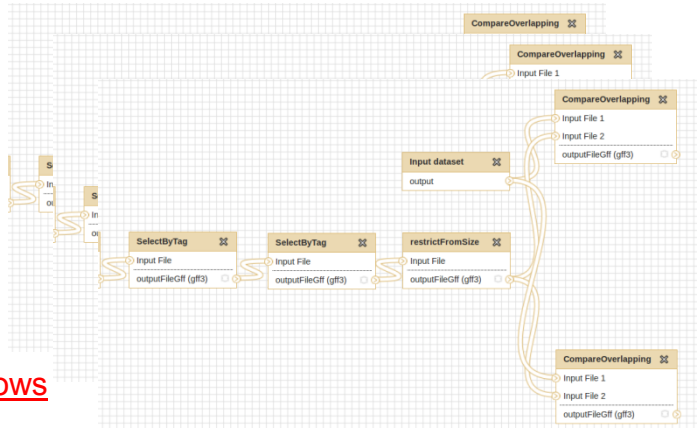
APLIBIO Projects

APLIBIO Projects



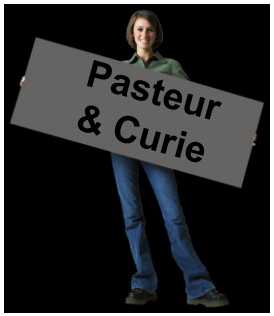
Pipelines (protocols)

VM?



workflows

APLIBIO Projects



Differential Analysis (DESeq)

Acknowledgment

Acknowledgment

Thank you for your attention !!!

URGI-MIG, Galaxy, 3/26/2012
yluo@versailles.inra.fr

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

Galaxy
Tool Intergration

- Easy to develop and add new tools but requires at least scripting skills
- Each tool is define into a xml file with a proper syntax.
  http://wiki.g2.bx.psu.edu/Admin/Tools/Tool Config Syntax
- Once the tool is created, you need to declare it in Galaxy by updating.
  $GALAXY_DIR/galaxy-dist/tool_conf.xml
- If you're using a job scheduler, you also need to declare the resources needed by your tool.
  Update $GALAXY_DIR/galaxy-dist/universe_wsgi.ini
- Restart Galaxy
- If you plan to make change in any tool xml file, you will have to reload the tool configuration from the admin tab (no need to restart Galaxy)

Example of tool creation: XML file

- **Tool and command definition:**

```
<tool id="upload_file" name="Upload file">
        <description>to current History</description>
        <command interpreter="bash">upload_file.sh -b $file -o $out1</command>
```

- **Inputs definition:**

```
  <inputs>
    <param name="name" type="text" label="File Name"/>
    <param name="extension" type="select" label="File type”>
      <option value="bam">Bam</option>
      <option value="txt">Text</option>
      <option value="fastq">Fastq</option>
      <option value="csfasta">Csfasta</option>
      <option value="qual">Qual</option>
      <option value="bed">Bed</option>
      <option value="gff">Gff</option>
      <option value="pdf">Pdf</option>
      <option value="vcf">VCF</option>
      <option value="sam">Sam</option>
      <option value="fasta">Fasta</option>
      <option value="xsq">Xsq</option>
    </param>
    <param name="file" type="text" size="30" label="Path to file"></param>
</inputs>
```

Example of tool creation: XML file

- **Outputs definition:**

```
<outputs>
        <data format="bam" name="out1" label="${name}.${extension}”>
                <change_format>
                        <when input="extension" value="txt" format="txt" />
                        <when input="extension" value="fastq" format="fastq" />
                        <when input="extension" value="csfasta" format="csfasta" />
                        <when input="extension" value="qual" format="qual" />
                        <when input="extension" value="bed" format="bed" />
                        <when input="extension" value="gff" format="gff" />
                        <when input="extension" value="vcf" format="vcf" />
                        <when input="extension" value="sam" format="sam" />
                        <when input="extension" value="fasta" format="fasta" />
                        <when input="extension" value="pdf" format="pdf" />
                        <when input="extension" value="xsq" format="xsq" />
                </change_format>
        </data>
</outputs>
<help>
</help>
</tool>
```

Example of tool creation: Execution file

(perl, python, shell, R, etc)

```
#!/bin/bash

while getopts "b:o:" optionName; do
case "$optionName" in

b) FILEORI="$OPTARG";;
o) FILEOUT="$OPTARG";;
esac
done

rm $FILEOUT

ln -s $FILEORI $FILEOUT
```

Example of tool creation: Tool declaration

Edit $GALAXY_DIR/galaxy-dist/tool_conf.xml:

```
<?xml version="1.0"?>
<toolbox>
        <section name="Get Data" id="getext">
                <tool file="data_source/upload.xml"/>
                <tool file="$PATH2NewTool/NewTool.xml" />
                <tool file="data_source/ucsc_tablebrowser.xml" />
                <tool file="data_source/ucsc_tablebrowser_test.xml" />
                <tool file="data_source/ucsc_tablebrowser_archaea.xml" />
                <tool file="data_source/bx_browser.xml" />
                <tool file="data_source/microbial_import.xml" />
                <tool file="data_source/biomart.xml" />
                <tool file="data_source/biomart_test.xml" />
        </section>
        ....
</toolbox>
```

Example of tool creation: Tool resources

Edit $GALAXY_DIR/galaxy-dist/universe_wsgi.ini:

```
# -- Database
# may use a SQLAlchemy connection string to specify an external database
# instead.  This string takes many options which are explained in detail in the
# config file documentation.

#database_connection = sqlite:///./database/universe.sqlite?isolation_level=IMMEDIATE
database_connection = postgres://userName:password@urgiDB.versailles.inra.fr:port/userDatabase
```

Then, you need to restart your Galaxy instance to start working with your tool