Biogenouest

ifb INSTITUT FRANÇAIS DE BIOINFORMATIQUE

OCEANOMICS

IDEALG seaweed for the future

ABiMS

14/12/2020

Galaxy

Initiation

Loraine Guéguen

Erwan Corre

Credits to Gildas Le Corguillé, Galaxy Training Network

v2.6

SORBONNE UNIVERSITÉ CRÉATEURS DE FUTURS DEPUIS 1257

cnrs

Learning objectives

- Learning objectives:
  - Familiarize yourself with the basics of Galaxy
  - Learn how to import data
  - Learn how to run tools
  - Learn how histories work
  - Learn how to create a workflow
  - Learn how to share your work
  - Understand and master dataset collections

Program

- Introduction
- Data import
- Tools
- Dataset
- History
- Workflow
- Share

Program

- Introduction
- Data import

- Tools
- Dataset
- History

Sequence files manipulation

- Workflow
- Share

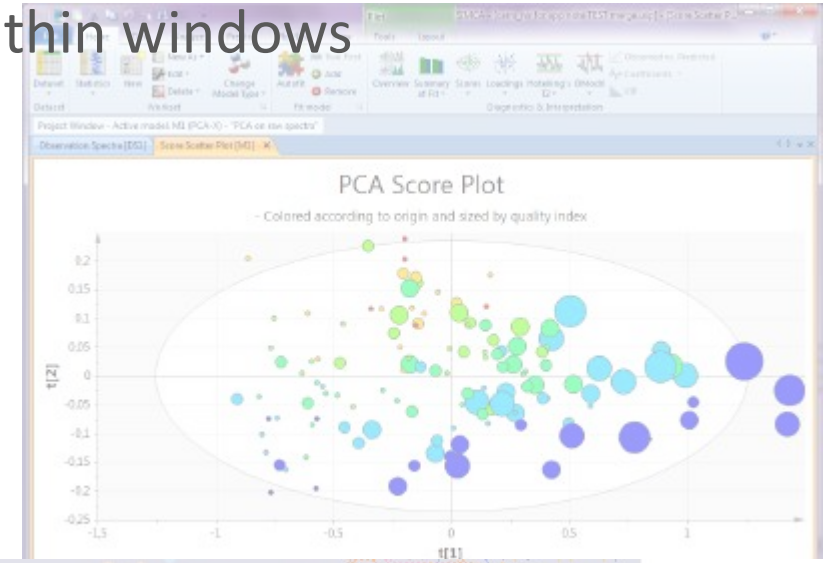
INTRODUCTION / PROBLEMATIC

Introduction / problematic

- In biomedical research, high-throughput technologies produce **large datasets**.
- How to perform analyses of these data **without bioinformatics skills** ?
    - Assemble transcript sequences de novo
    - Determine the gene expression
    - Build a phylogenetic tree
    - Predict subcellular targeting for proteins
    - Identify and quantify metabolites detected by LC-MS
    - …

Introduction / problematic

- **Graphical interface** click-button tools within windows
  - + very ergonomic
  - – too ergonomic → lack of flexibility
  - – few
  - – paying for it!



- Tools available on the **internet**
  - + very ergonomic
  - – too ergonomic → lack of flexibility
  - – A small part of the available tools
  - – the submission size /storage is often limited
  - – must not be paranoid

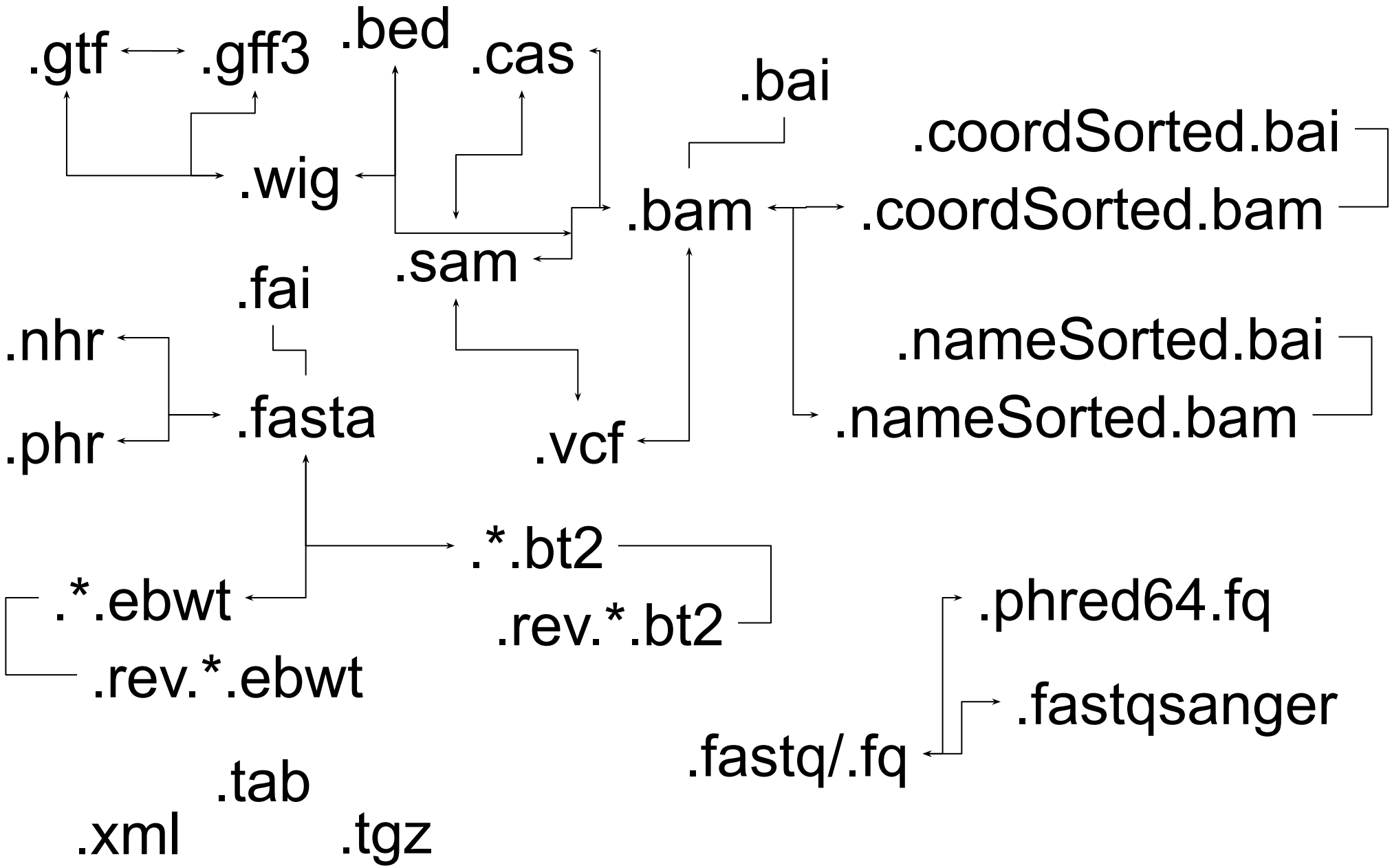
Introduction / problematic

- **Command line** tools
  - \+ represent almost the majority of scientific tools
  - \+ good parameters completeness
  - \+ can be executed on high performance computers
  - \+ automatable, workflowsable, ...
  - – minimum linux knowledge is required
  - – cruel lack of ergonomics

```
login@sbr4-1042:~$ ssh -Y login@bioinfo.sb-roscoff.fr
[...]
[login@n0 ~]$ cdprojet
[login@n0 login]$ cd 13-07-29-panda/tmp/mapping
[login@n0 mapping]$ cat tophat.qsub
#!/bin/bash
#$ -S /bin/bash
#$ -M login@sb-roscoff.fr
#$ -m bea
#$ -V
#$ -cwd
#$ -o qsub.out
#$ -e qsub.err

tophat2 panda_v121029 ../input/IllR1-1.fq ../input/IllR1-2.fq
-GTF ../input/panda_v121029.gtf --b2-sensitive -r 100
-num-threads 8

[login@n0 mapping]$ qsub -q long.q -pe thread 8 tophat.qsub
Your job 5338969 ("tophat.qsub") has been submitted
[login@n0 mapping]$ ls
accepted_hits.bam  junctions.bed    qsub.err  unmapped.bam
deletions.bed      logs             qsub.out
insertions.bed     prep_reads.info  tmp
[login@n0 mapping]$ cd
```

```
login@sbr4-1042:~$ ssh -Y login@bioinfo.sb-roscoff.fr
[...]
[login@n0 ~]$ cdproiet
[login@n0 login]$
[login@n0 mapping
#!/bin/bash
#$ -S /bin/bash
#$ -M login@sb-ro
#$ -m bea
#$ -V
#$ -cwd
#$ -o qsub.out
#$ -e qsub.err

tophat2 panda_v12                                          .R1-2.fq
-GTF ../input/par
-num-threads 8

[login@n0 mapping                                          :.qsub
Your job 5338969
[login@n0 mapping
accepted_hits.bam  junctions.bed    qsub.err  unmapped.bam
deletions.bed      logs             qsub.out
insertions.bed     prep_reads.info  tmp
[login@n0 mapping]$ cd
```

Introduction / problematic



Introduction / problematic

Select your level:

Introduction / problematic

Level 1

« I want to know the gene expression »

Introduction / problematic

Level 2

« I want to map my reads on a reference genome and count them »

Introduction / problematic

Level 3

« I want to launch the tools tophat2 and cufflinks.

I have fastq files and my genome in fasta and gtf. »

Introduction / problematic

Level 4

« I want 1TB for my project. I will launch tophat2 through SSH on the cluster in multi-thread mode.

Next I want to submit the bam file to my genome with cufflinks.

Except that, I will manage :P”

Introduction / problematic

Level 5

« I have a bunch of cool tools!
But I'm the only one who can launch them.

Comments? »

Galaxy

- Web-based platform for computational biomedical research (analysis and data integration)
  - Developed at Penn State, Johns Hopkins, OHSU and Cleveland Clinic with substantial outside contributions
  - Open source under Academic Free License
- More than 6,500 citations
- More than 125 public Galaxy resources
  - 100+ public servers, many more non-public
  - Both general-purpose and domain-specific

INTRODUCTION / GALAXY

Introduction / Galaxy

Why Galaxy ?

- Accessibility
- Reproductibility
- Transparency

Introduction / Galaxy

- Galaxy it's …
  - A web-based interface
  - No need to execute a command line through a terminal
  - Programming or scripting skills are not required
  - Submission of jobs is transparent through a high performance computer cluster
  - Secure histories and data manager
  - A sharing system for data and protocols
  - Tool-boxes of several bioinformatics fields
    - NGS
    - Chemistry
    - Metabolomics
    - Image analysis
    - Statistics
    - Etc ...

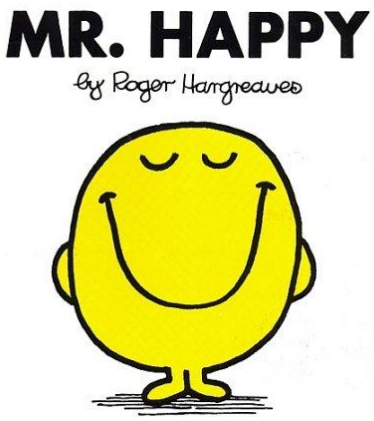
Introduction / Galaxy

MR. GEEK

```
[login@n0 ~]$ cdprojet
[login@n0 login]$ cd 13-07-29-panda/tmp/mapping
[login@n0 mapping]$ cat tophat.qsub
#!/bin/bash
#$ -S /bin/bash
#$ -M login@sb-roscoff.fr
#$ -m bea
#$ -V
#$ -cwd
#$ -o qsub.out
#$ -e qsub.err

tophat2 panda_v121029 ../input/IllR1-1.fq ../input/IllR1-2.fq
-GTF ../input/panda_v121029.gtf --b2-sensitive -r 100
-num-threads 8

[login@n0 mapping]$ qsub -q long.q -pe thread 8 tophat.qsub
Your job 5338969 ("tophat.qsub") has been submitted
```

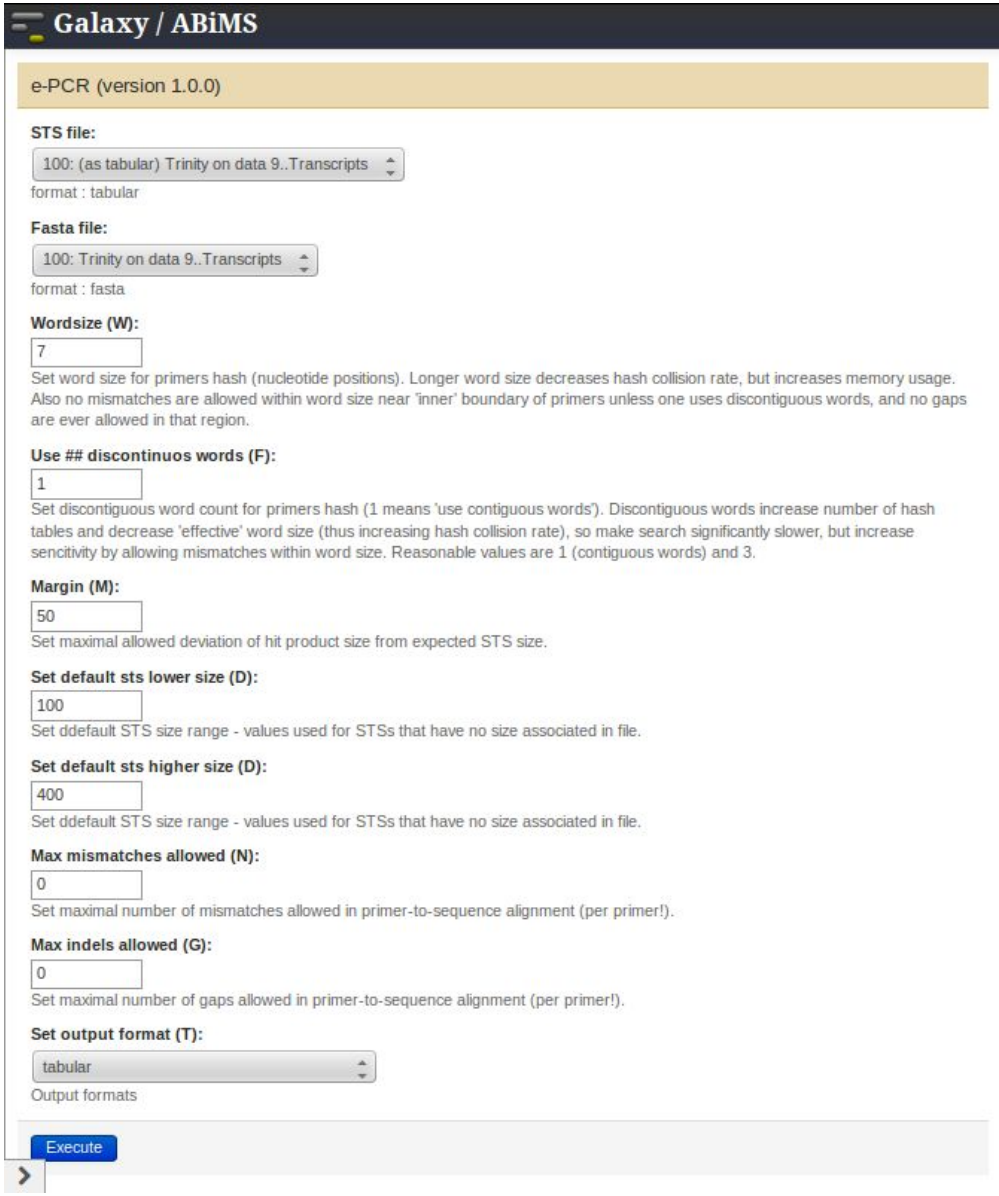
Introduction / Galaxy

```
[login@n0 ~]$ cdprojet
[login@n0 login]$ cd 13-07-29-panda/tmp/mapping
[login@n0 mapping]$ cat tophat.qsub
#!/bin/bash
#$ -S /bin/bash
#$ -M login@sb-roscoff.fr
#$ -m bea
#$ -V
#$ -cwd
#$ -o qsub.out
#$ -e qsub.err

tophat2 panda_v121029 ../input/Ill1R1-1.fq ../input/Ill1R1-2.fq
-GTF ../input/panda_v121029.gtf --b2-sensitive -r 100
-num-threads 8

[login@n0 mapping]$ qsub -q long.q -pe thread 8 tophat.qsub
Your job 5338969 ("tophat.qsub") has been submitted
```



Introduction / Galaxy

```
[lecorguille@n0 ~]$ e-PCR --help
e-PCR: invalid option -- -
usage: [-hV] [posix-options] stsfile [fasta ...]
[compat-options]
where posix-options are:
     -m ##     Margin (default 50)
     -w ##     Wordsize  (default 7)
     -n ##     Max mismatches allowed (default 0)
     -g ##     Max indels allowed (default 0)
     -f ##     Use ## discontiguos words, slow if
                ##>1
     -o ##     Set output file
     -t ##     Set output format:
          1 - classic, range (pos1..pos2)
          2 - classic, midpoint
          3 - tabular
          4 - tabular with alignment in comments
               (slow)
     -d##-##   Set default size range
               (default 100-350)
     -p +-     Turn hits postprocess on/off
     -v ##     Verbosity flags
     -a a|f    Use presize alignmens (only if
               gaps>0), slow
           a - Allways or f - as Fallback
     -x +-     Use 5'-end lowercase masking of
               primers (default -)
     -u +-     Uppercase all primers (default -)

[...]
```

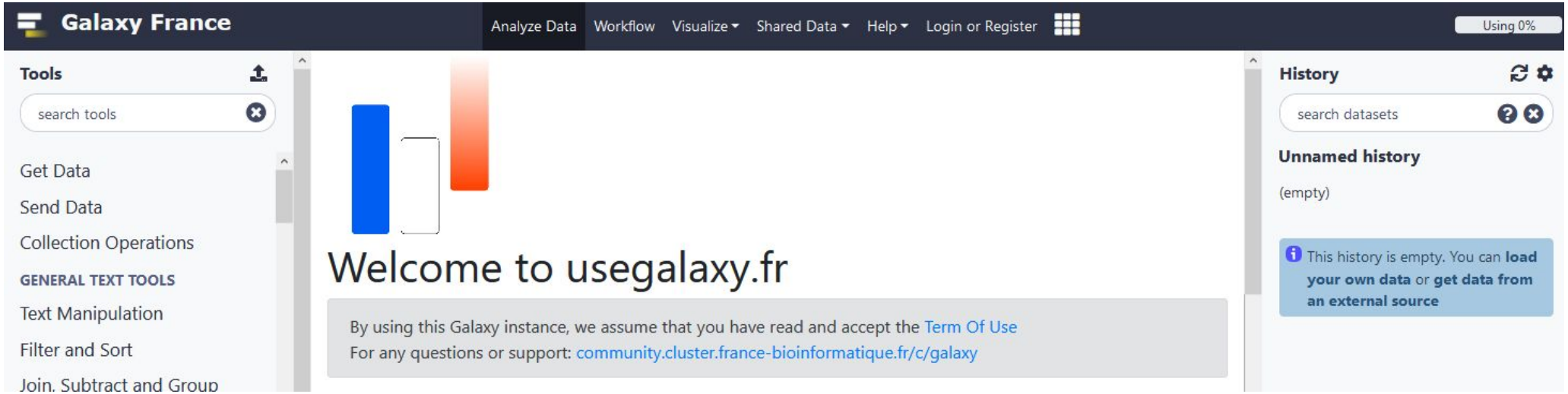


Introduction / Galaxy

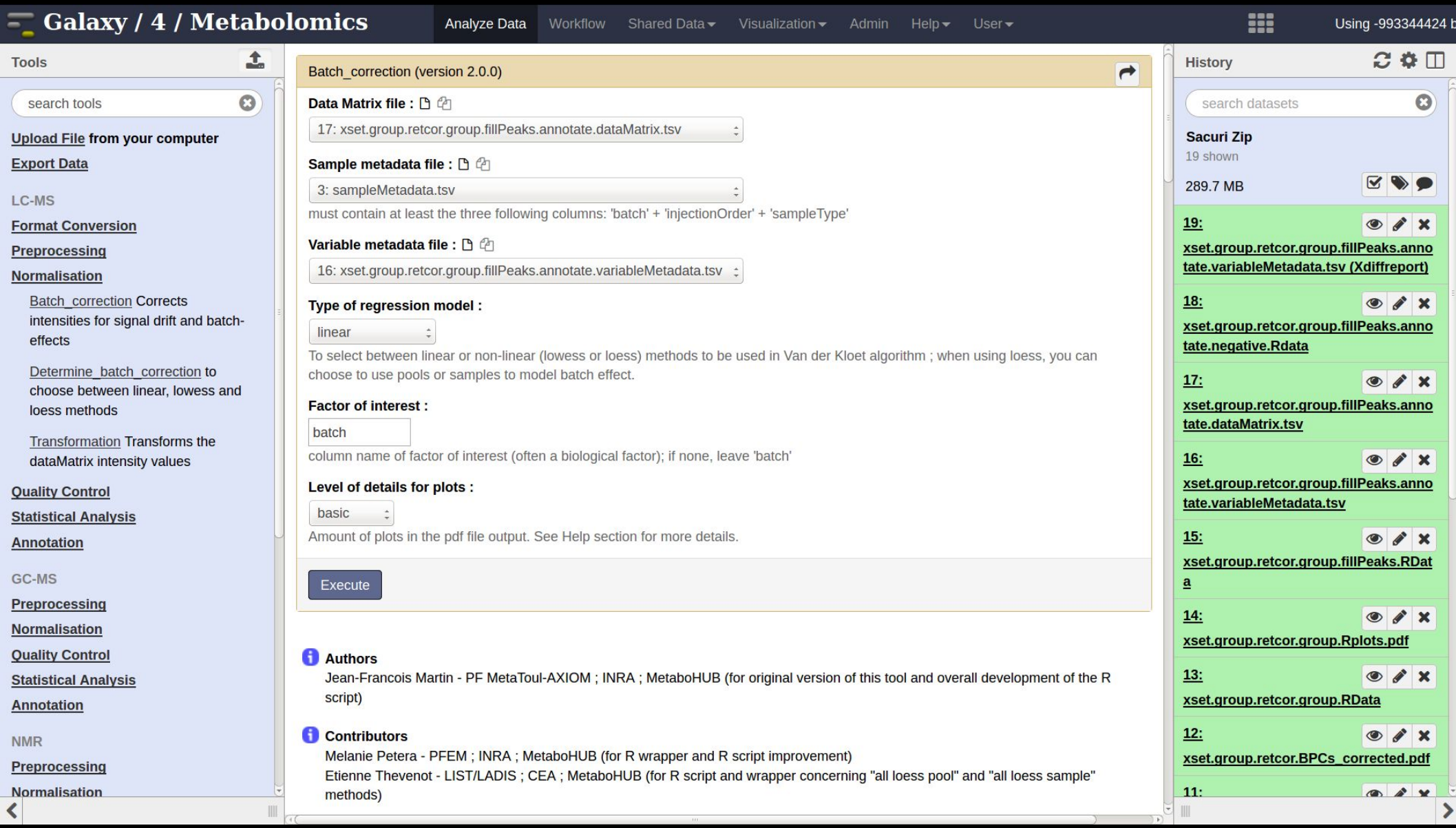
Galaxy ABiMS --> Galaxy France

https://galaxy.sb-roscoff.fr/

https://usegalaxy.fr/

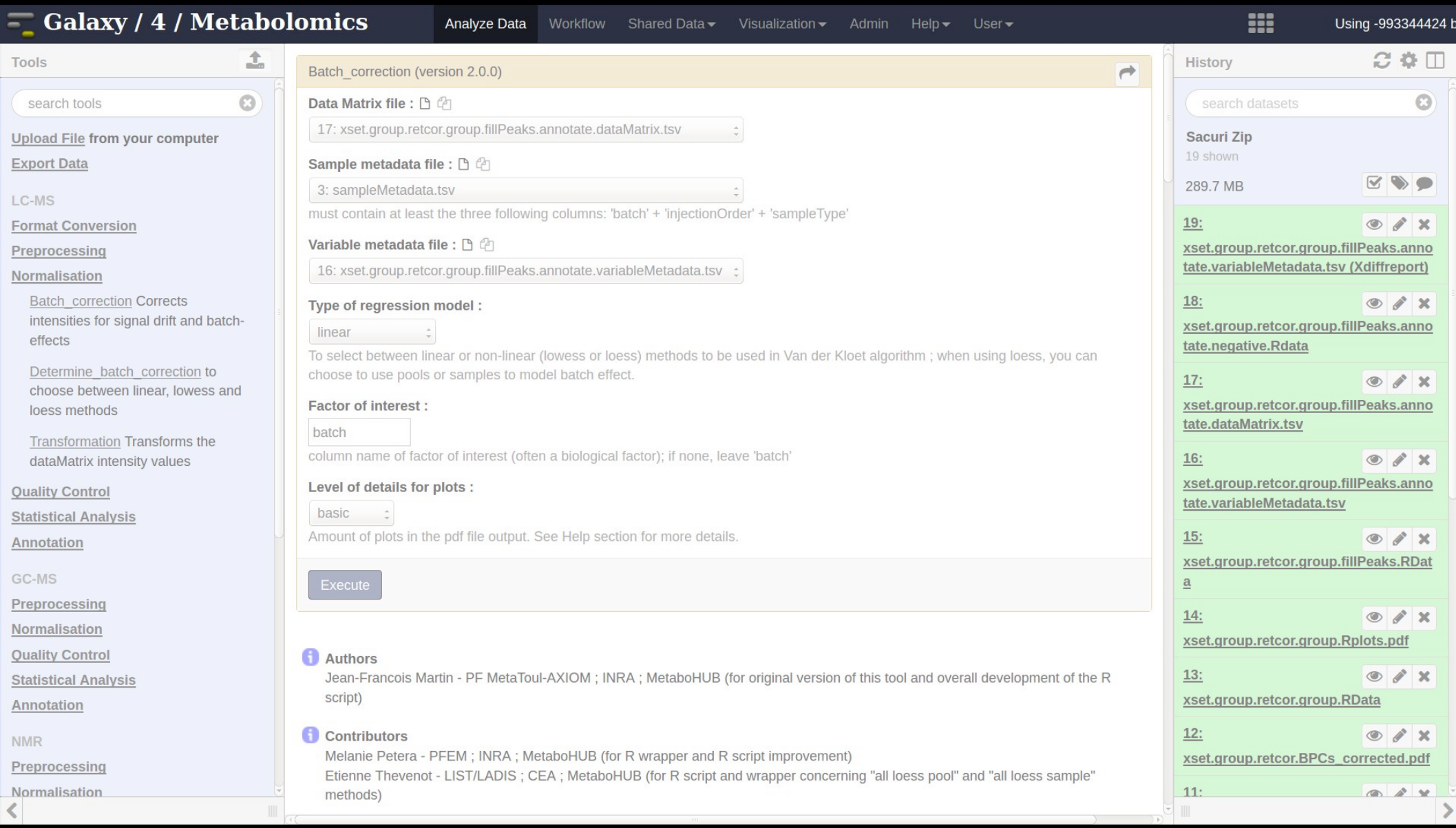


Galaxy interface



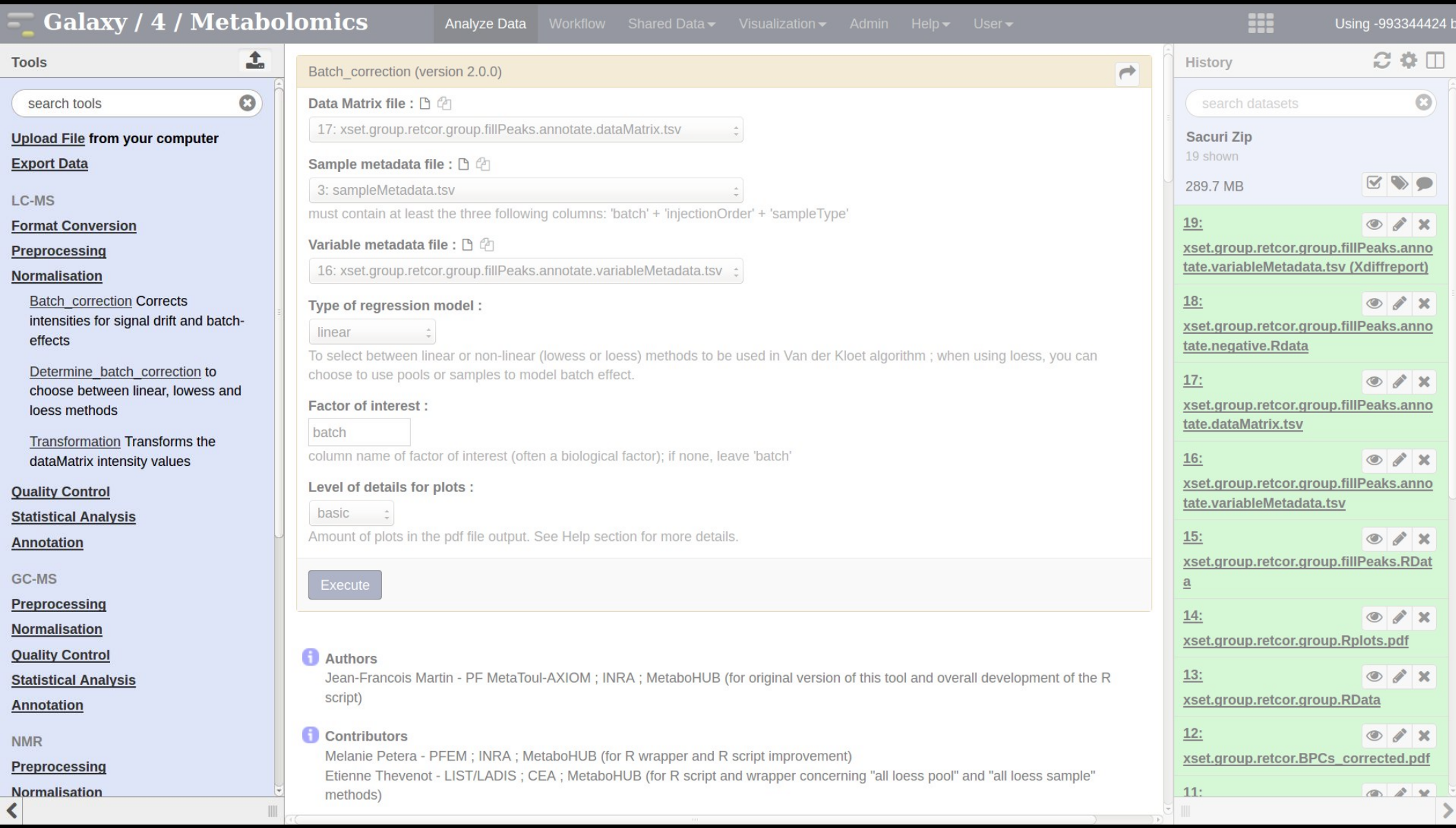
Galaxy interface

Top menu



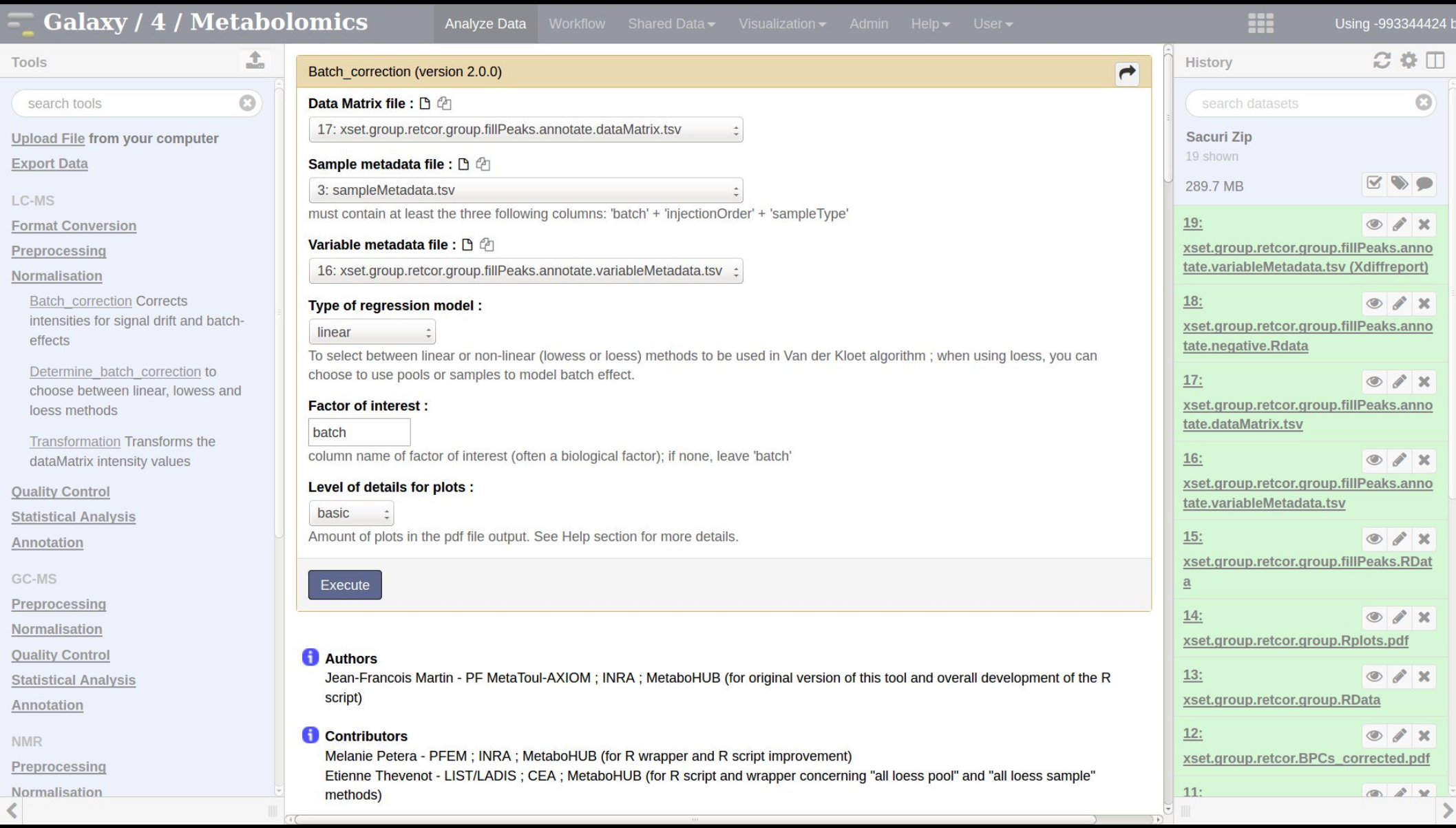
Galaxy interface

Tool list



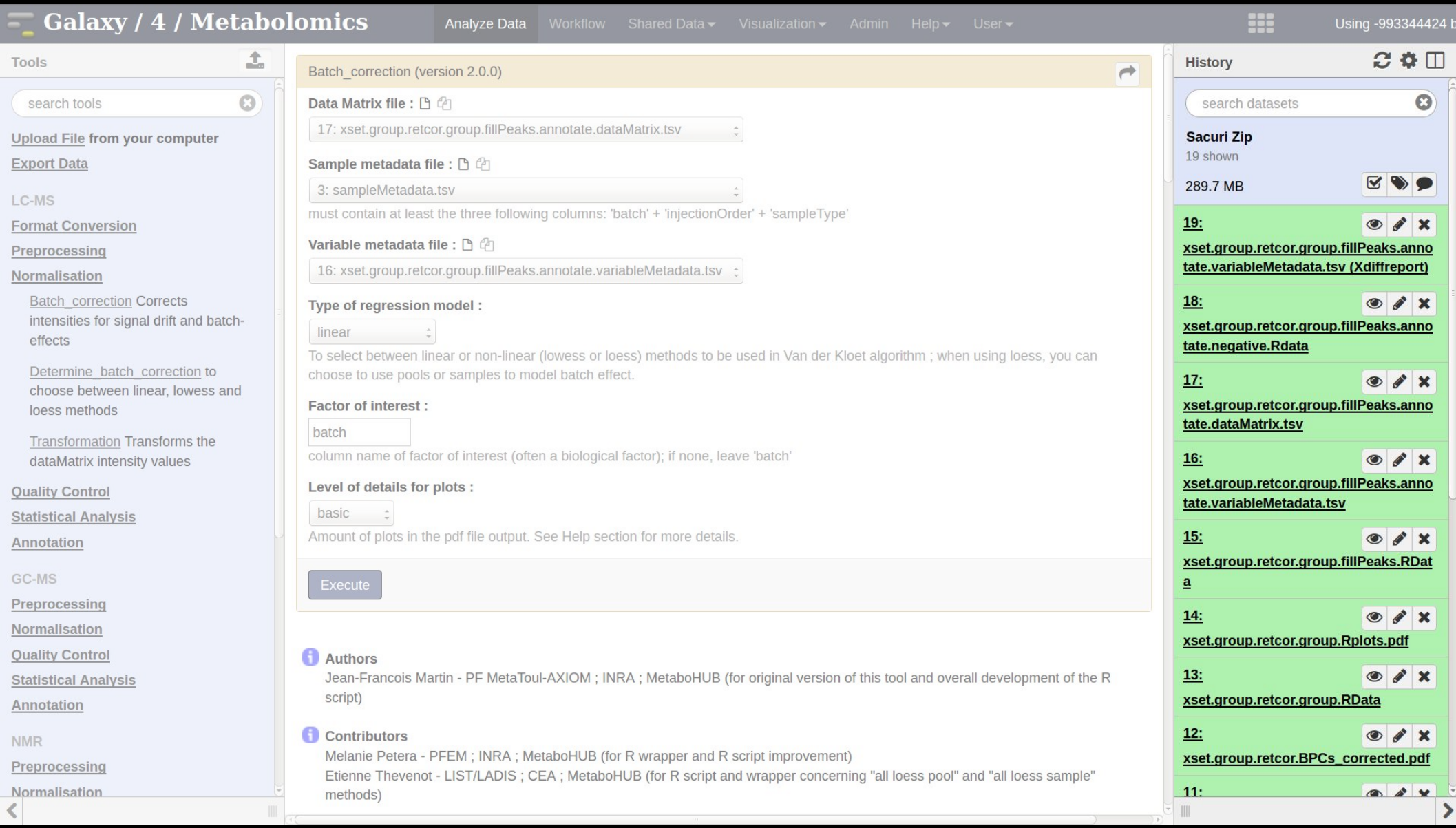
Galaxy interface

Web forms / dataset visualization / miscellaneous information



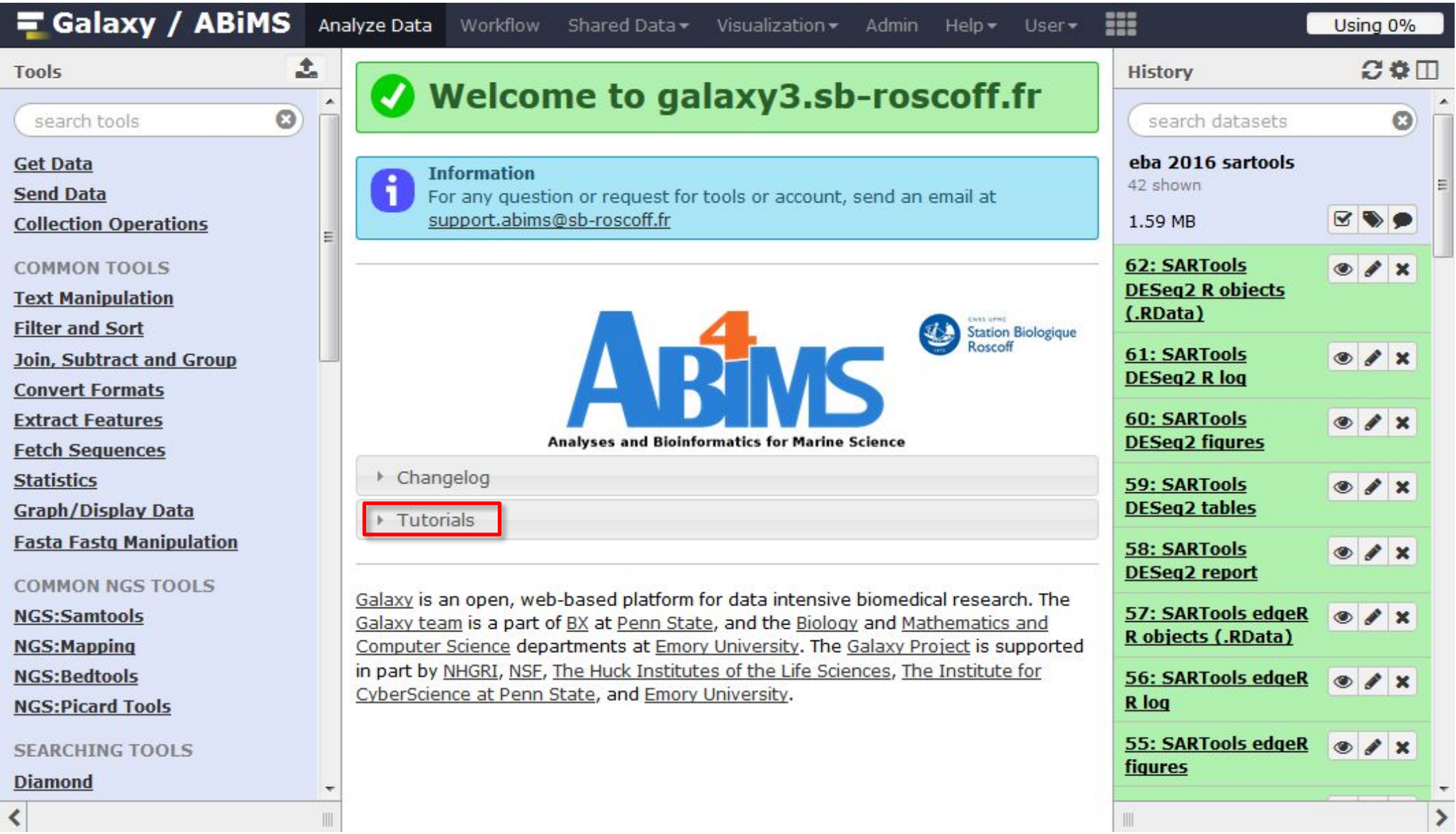
Galaxy interface

History

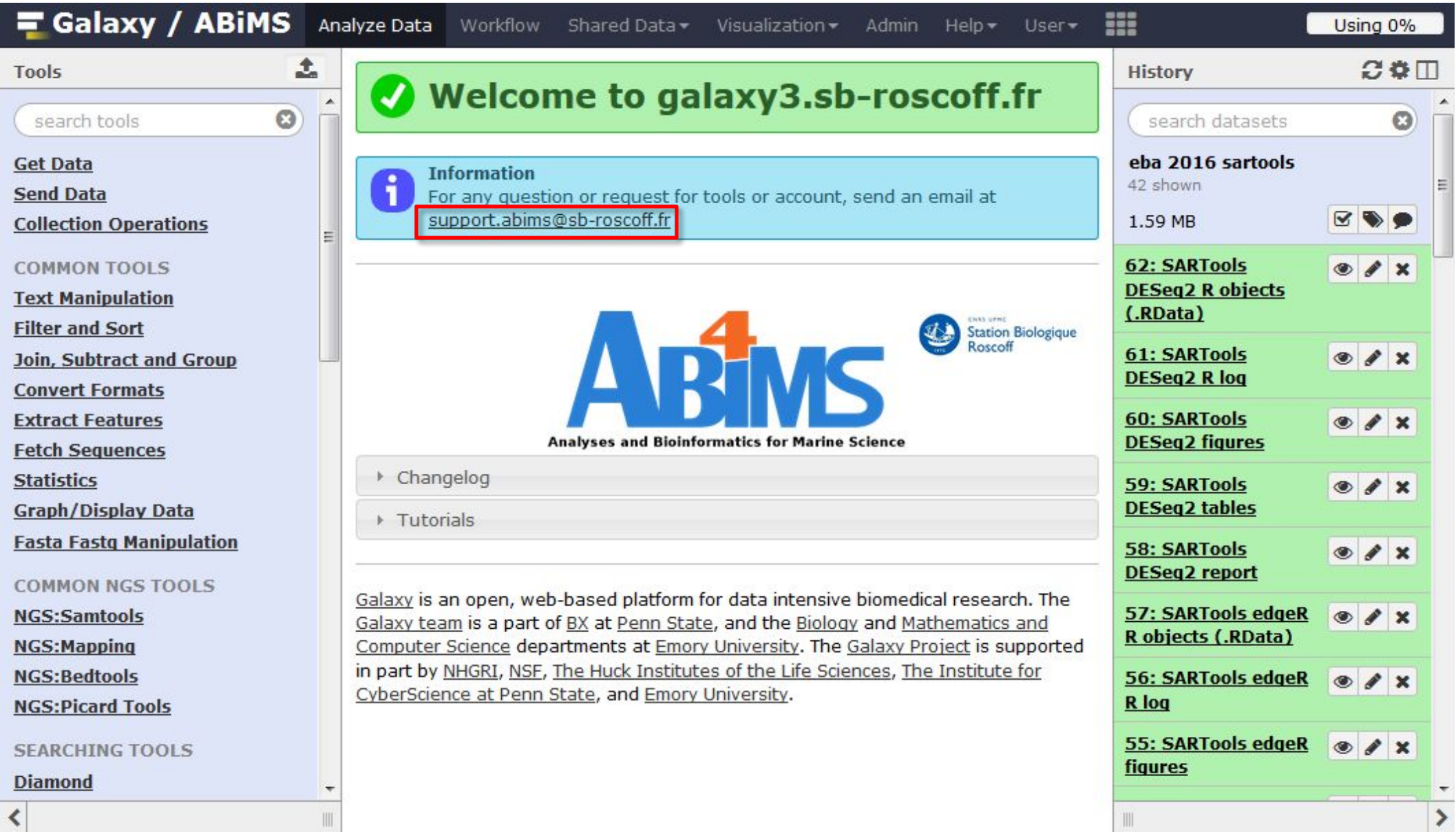


GET HELP

Get help


Galaxy / ABiMS
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 0%
Tools
search tools
Get Data
Send Data
Collection Operations
COMMON TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Statistics
Graph/Display Data
Fasta Fastq Manipulation
COMMON NGS TOOLS
NGS:Samtools
NGS:Mapping
NGS:Bedtools
NGS:Picard Tools
SEARCHING TOOLS
Diamond
Welcome to galaxy3.sb-roscoff.fr
Information
For any question or request for tools or account, send an email at support.abims@sb-roscoff.fr
Analyses and Bioinformatics for Marine Science
Station Biologique Roscoff
Changelog
Tutorials
Galaxy is an open, web-based platform for data intensive biomedical research. 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 NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.
History
search datasets
eba 2016 sartools
42 shown
1.59 MB
62: SARTools DESeq2 R objects (.RData)
61: SARTools DESeq2 R log
60: SARTools DESeq2 figures
59: SARTools DESeq2 tables
58: SARTools DESeq2 report
57: SARTools edgeR R objects (.RData)
56: SARTools edgeR R log
55: SARTools edgeR figures

Get help


Galaxy / ABiMS
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 0%
Tools
search tools
Get Data
Send Data
Collection Operations
COMMON TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Statistics
Graph/Display Data
Fasta Fastq Manipulation
COMMON NGS TOOLS
NGS:Samtools
NGS:Mapping
NGS:Bedtools
NGS:Picard Tools
SEARCHING TOOLS
Diamond
Welcome to galaxy3.sb-roscoff.fr
Information
For any question or request for tools or account, send an email at
support.abims@sb-roscoff.fr
Analyses and Bioinformatics for Marine Science
Station Biologique Roscoff
Changelog
Tutorials
Galaxy is an open, web-based platform for data intensive biomedical research. 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 NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.
History
search datasets
eba 2016 sartools
42 shown
1.59 MB
62: SARTools DESeq2 R objects (.RData)
61: SARTools DESeq2 R log
60: SARTools DESeq2 figures
59: SARTools DESeq2 tables
58: SARTools DESeq2 report
57: SARTools edgeR R objects (.RData)
56: SARTools edgeR R log
55: SARTools edgeR figures

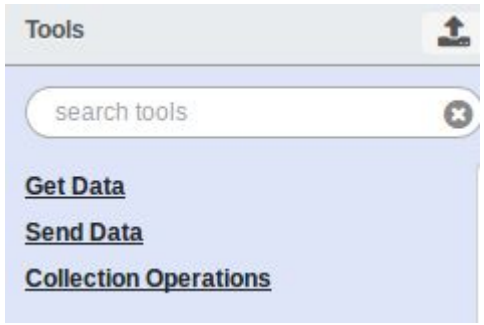
DATA IMPORT

Data import

- Copy/paste from a file
- Upload data from your local computer
- Upload data from internet using URL
- Upload data from FTP
- Upload data from online databases: EBI ENA, UCSC, BioMart, etc.
- Import from Shared Data (libraries, histories, workflows)

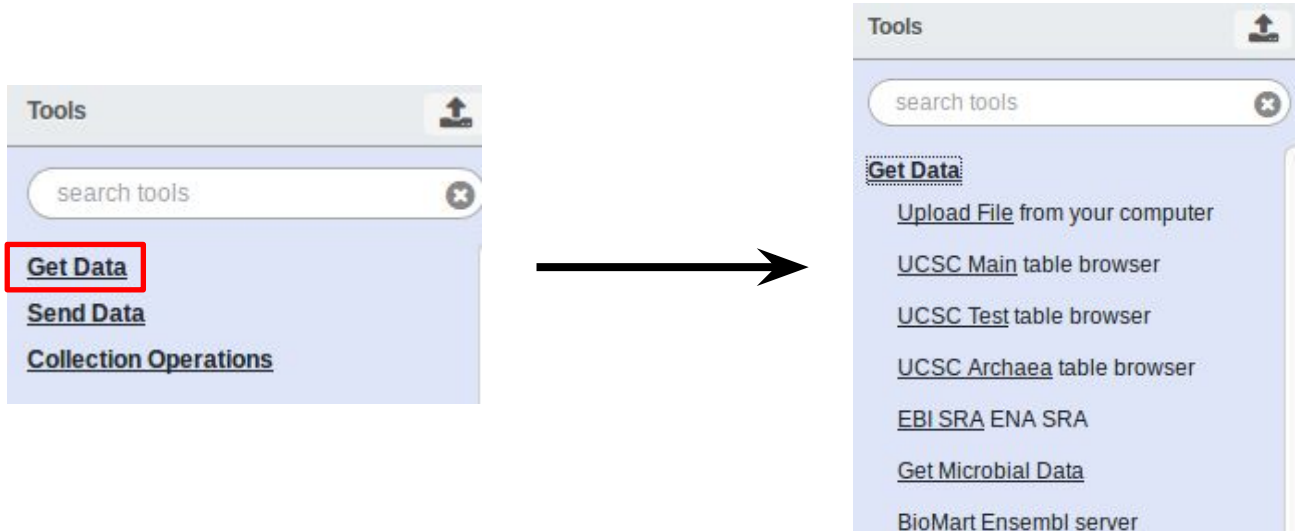
Data import

- Copy/paste from a file
- Upload data from your local computer
- Upload data from internet using URL
- Upload data from FTP
- Upload data from online databases: EBI ENA, UCSC, BioMart, etc.
- Import from Shared Data (libraries, histories, workflows)



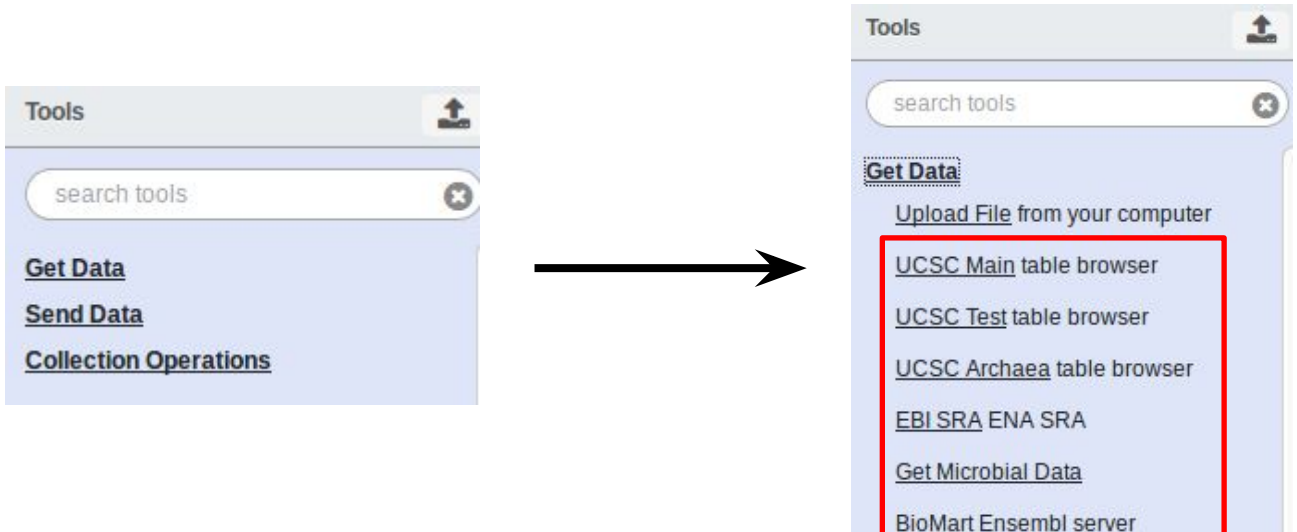
Data import

- Copy/paste from a file
- Upload data from your local computer
- Upload data from internet using URL
- Upload data from FTP
- Upload data from online databases: EBI ENA, UCSC, BioMart, etc.
- Import from Shared Data (libraries, histories, workflows)



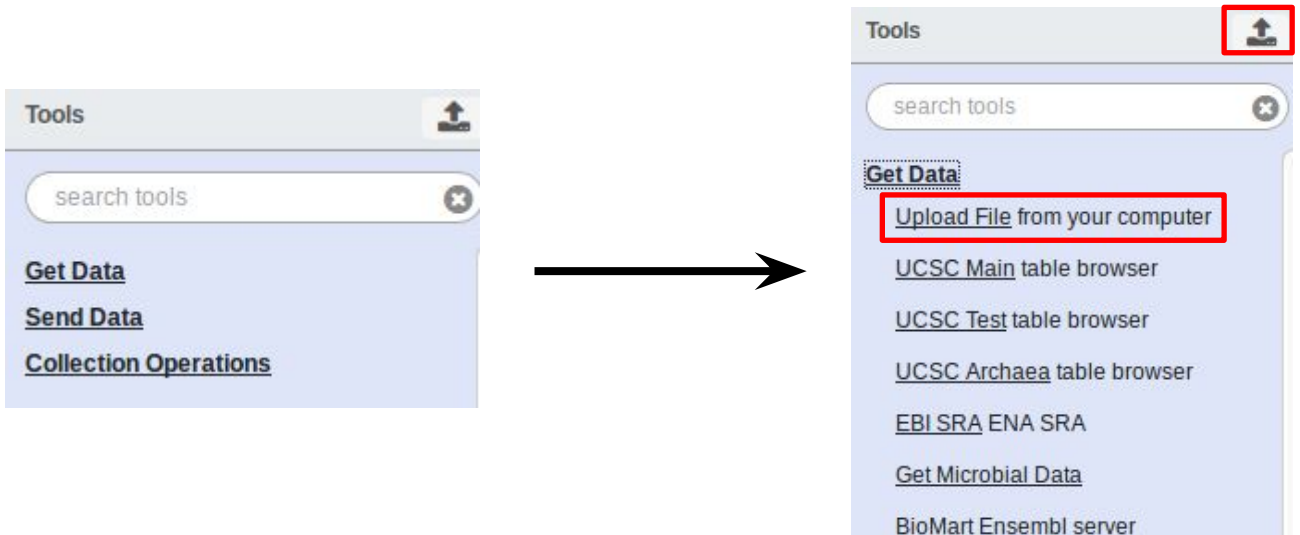
Data import

- Copy/paste from a file
- Upload data from your local computer
- Upload data from internet using URL
- Upload data from FTP
- Upload data from online databases: EBI ENA, UCSC, BioMart, etc.
- Import from Shared Data (libraries, histories, workflows)

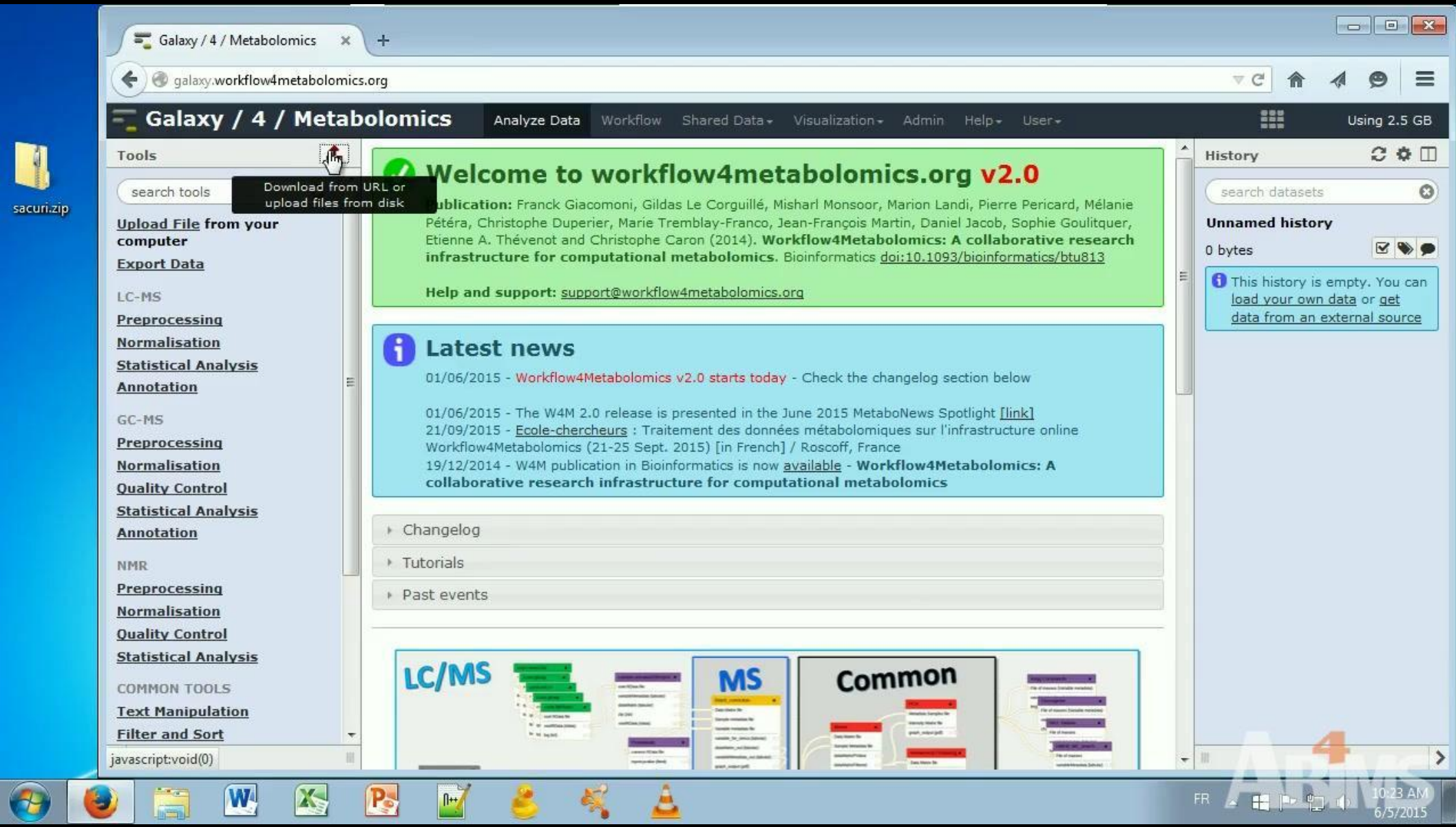


Data import

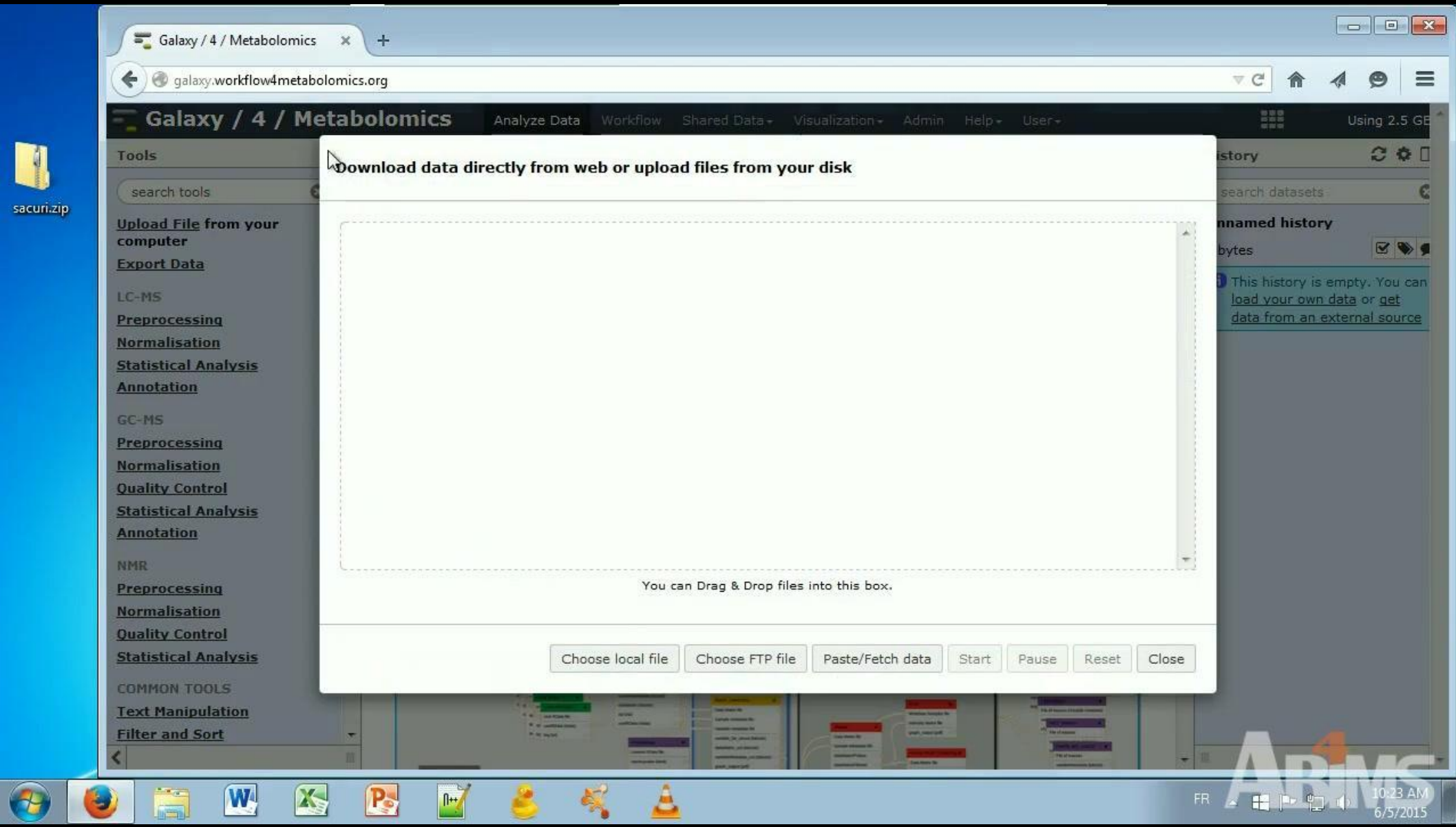
- Copy/paste from a file
- Upload data from your local computer
- Upload data from internet using URL
- Upload data from FTP
- Upload data from online databases: EBI ENA, UCSC, BioMart, etc.
- Import from Shared Data (libraries, histories, workflows)



Data import

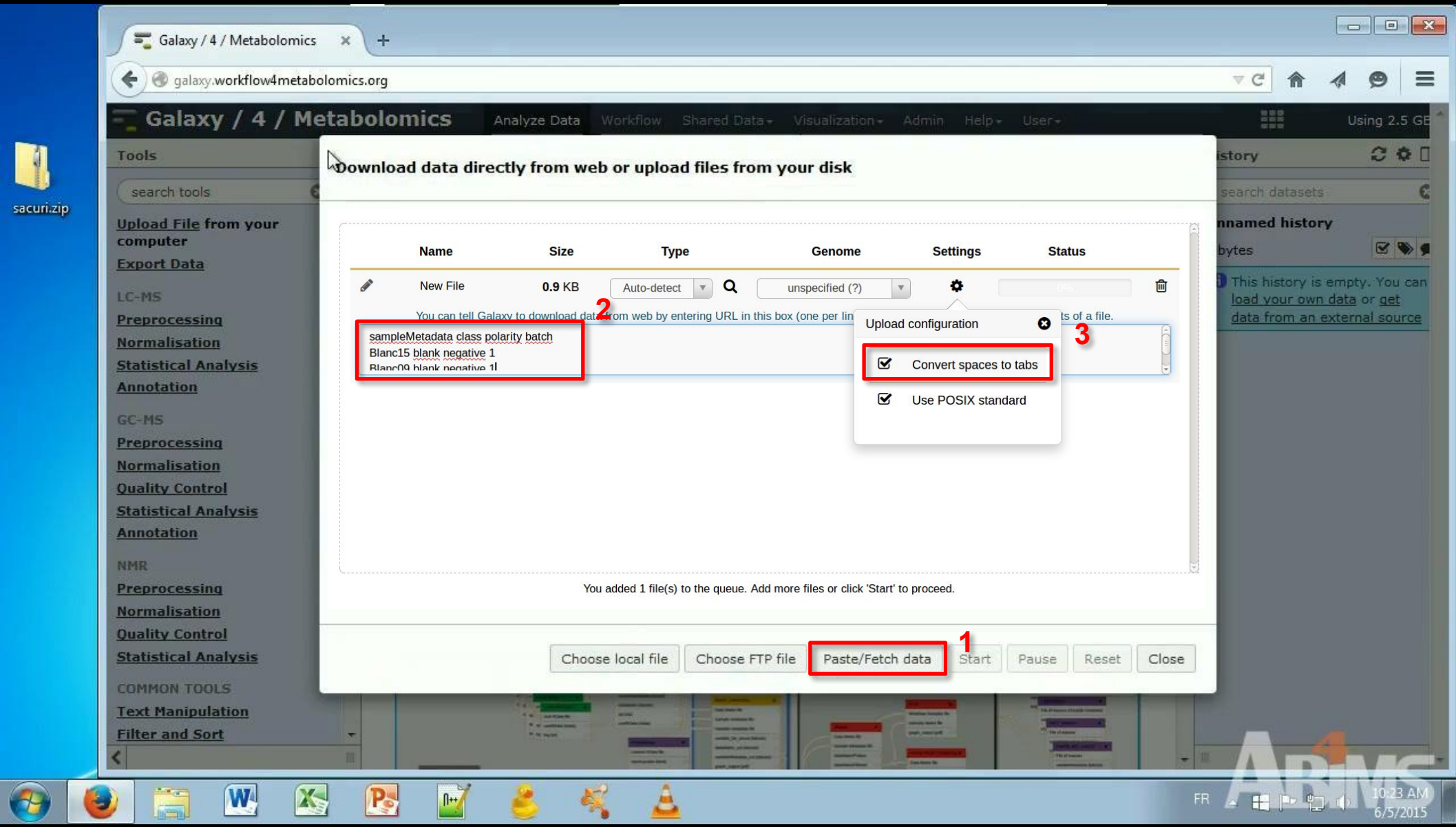

Galaxy / 4 / Metabolomics
galaxy.workflow4metabolomics.org
Galaxy / 4 / Metabolomics
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 2.5 GB
Tools
search tools
Download from URL or upload files from disk
Upload File from your computer
Export Data
LC-MS
Preprocessing
Normalisation
Statistical Analysis
Annotation
GC-MS
Preprocessing
Normalisation
Quality Control
Statistical Analysis
Annotation
NMR
Preprocessing
Normalisation
Quality Control
Statistical Analysis
COMMON TOOLS
Text Manipulation
Filter and Sort
javascript:void(0)
Welcome to workflow4metabolomics.org v2.0
ublication: Franck Giacomoni, Gildas Le Corguillé, Misharl Monsoor, Marion Landi, Pierre Pericard, Mélanie Pétéra, Christophe Duperier, Marie Tremblay-Franco, Jean-François Martin, Daniel Jacob, Sophie Goulitquer, Etienne A. Thévenot and Christophe Caron (2014). Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics. Bioinformatics doi:10.1093/bioinformatics/btu813
Help and support: support@workflow4metabolomics.org
Latest news
01/06/2015 - Workflow4Metabolomics v2.0 starts today - Check the changelog section below
01/06/2015 - The W4M 2.0 release is presented in the June 2015 MetaboNews Spotlight [link]
21/09/2015 - Ecole-chercheurs : Traitement des données métabolomiques sur l'infrastructure online Workflow4Metabolomics (21-25 Sept. 2015) [in French] / Roscoff, France
19/12/2014 - W4M publication in Bioinformatics is now available - Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics
Changelog
Tutorials
Past events
LC/MS
MS
Common
History
search datasets
Unnamed history
0 bytes
This history is empty. You can load your own data or get data from an external source
sacuri.zip
FR
10:23 AM
6/5/2015

Data import


Galaxy / 4 / Metabolomics
galaxy.workflow4metabolomics.org
Galaxy / 4 / Metabolomics
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 2.5 GB
Tools
search tools
Upload File from your computer
Export Data
LC-MS
Preprocessing
Normalisation
Statistical Analysis
Annotation
GC-MS
Preprocessing
Normalisation
Quality Control
Statistical Analysis
Annotation
NMR
Preprocessing
Normalisation
Quality Control
Statistical Analysis
COMMON TOOLS
Text Manipulation
Filter and Sort
Download data directly from web or upload files from your disk
You can Drag & Drop files into this box.
Choose local file
Choose FTP file
Paste/Fetch data
Start
Pause
Reset
Close
search datasets
Unnamed history
0 bytes
This history is empty. You can load your own data or get data from an external source
sacuri.zip
FR
10:23 AM
6/5/2015

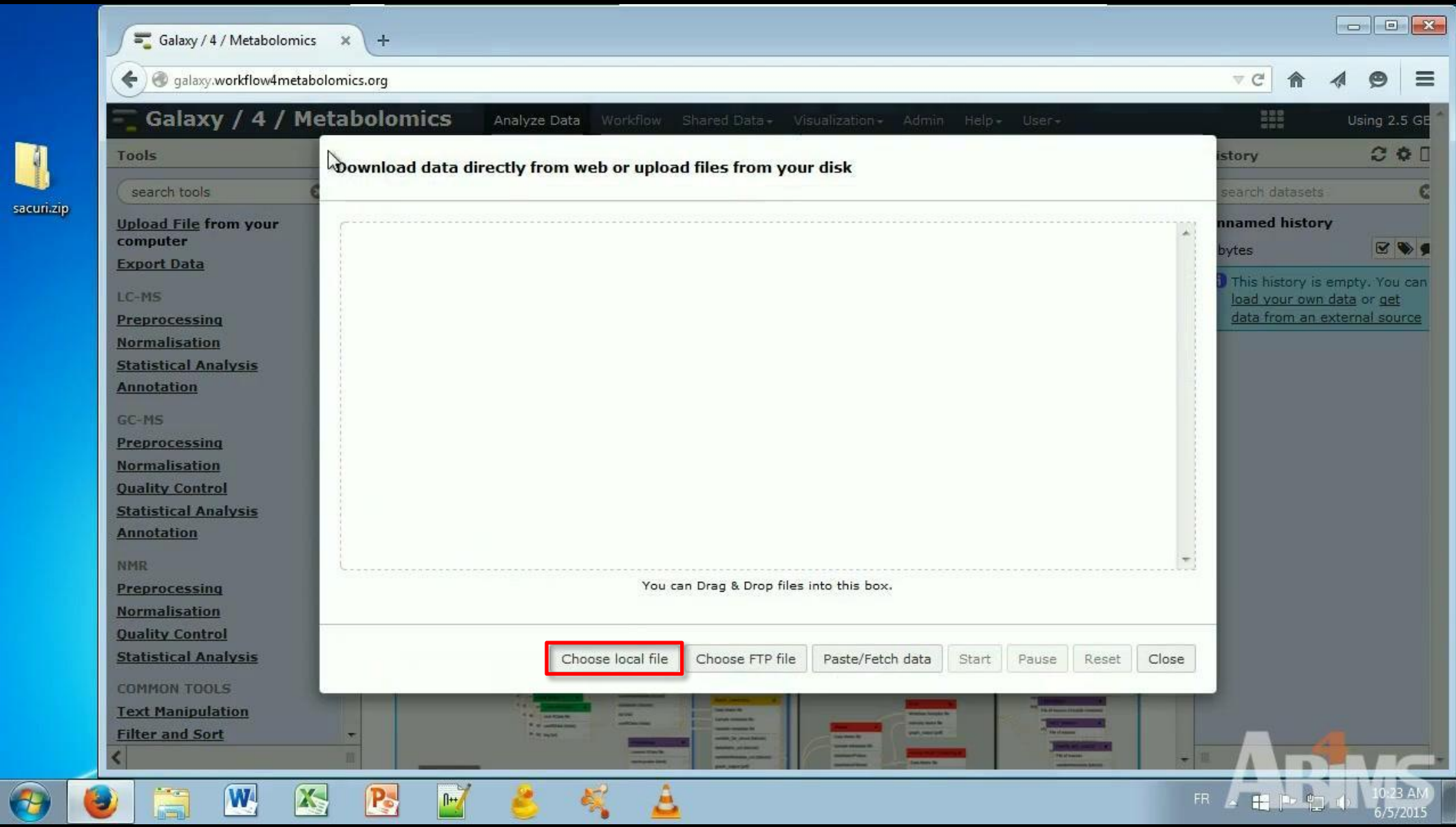
Data import

Copy / Paste data



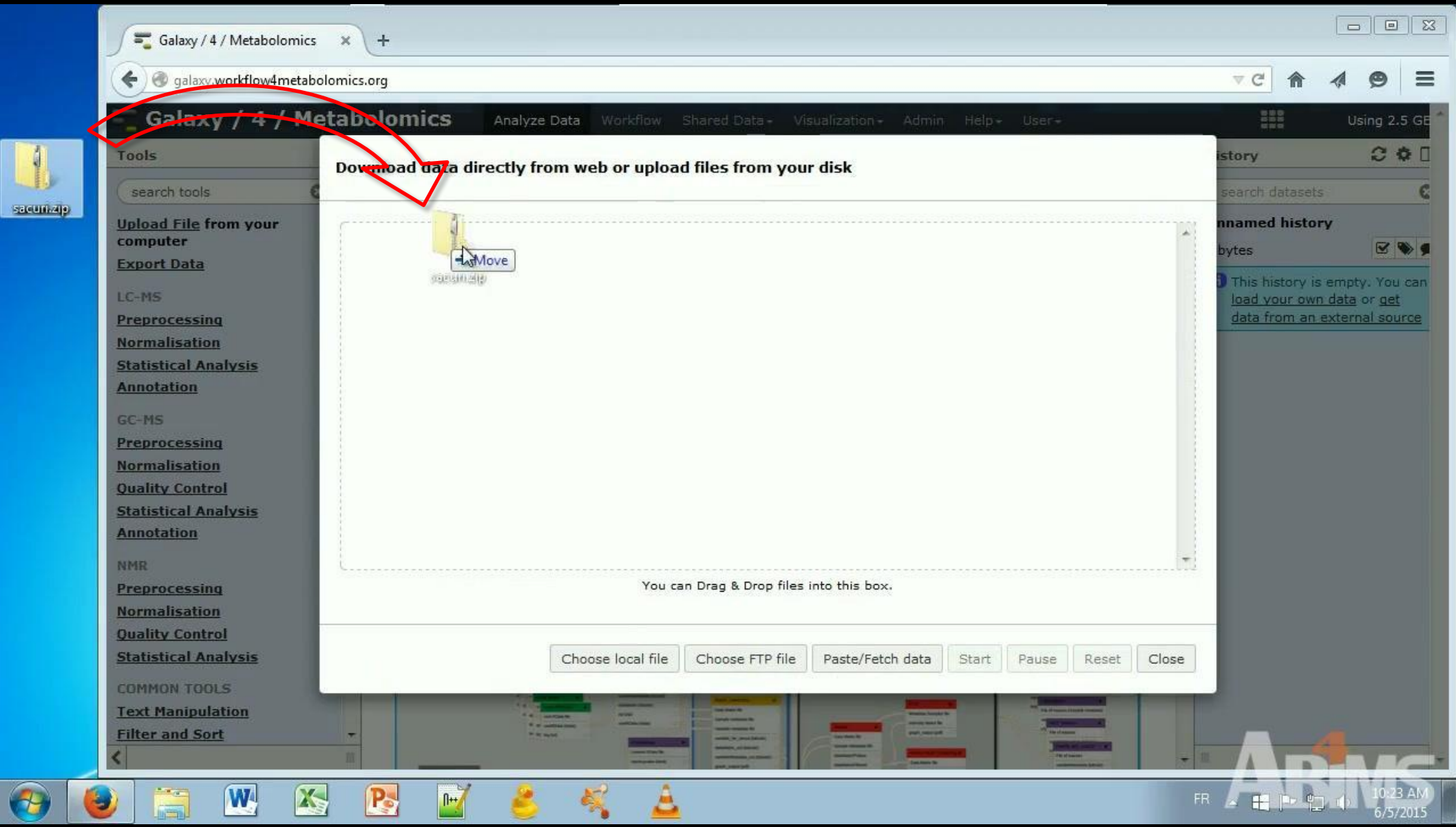
Data import

From local files



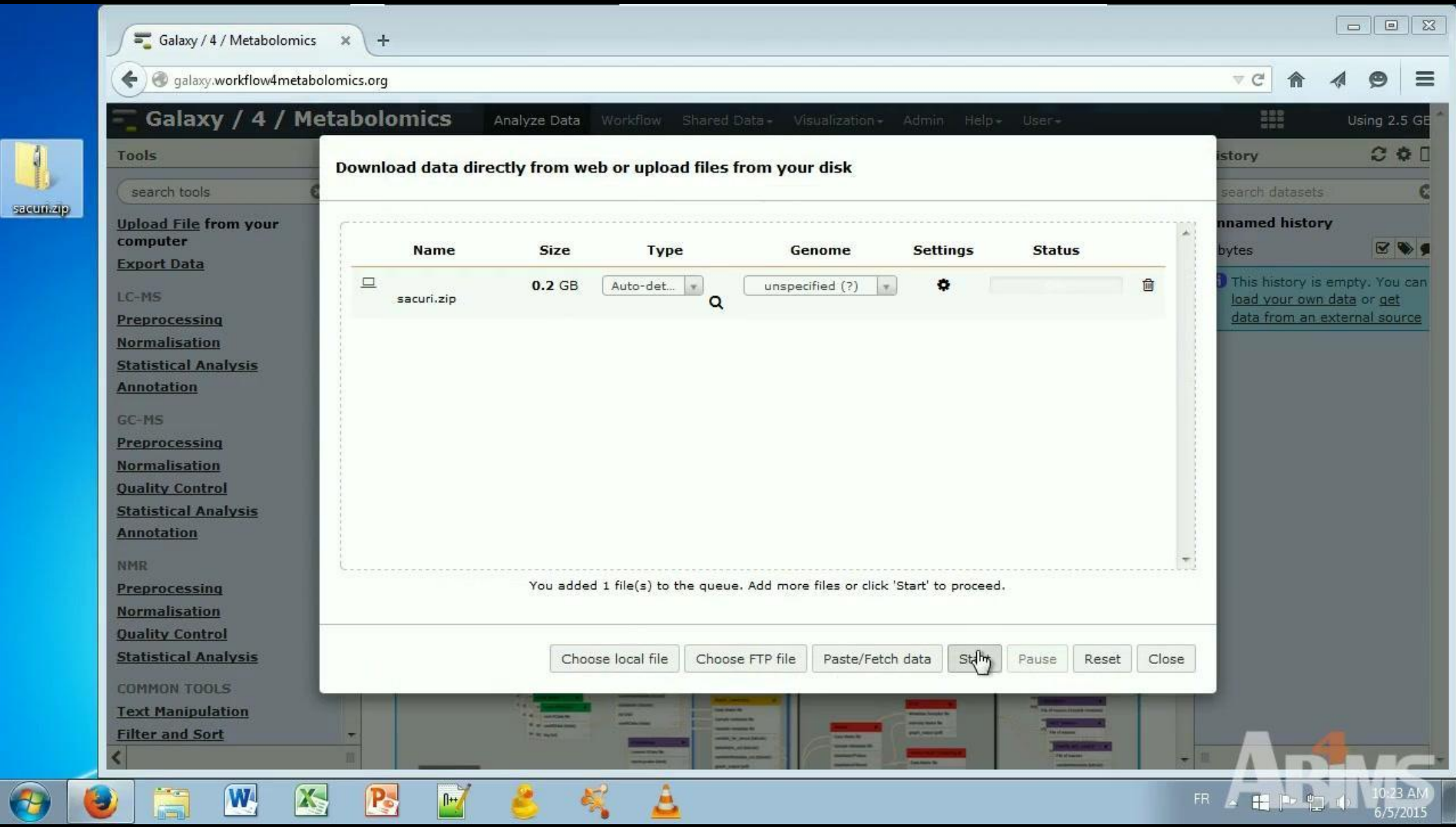
Data import

From local files



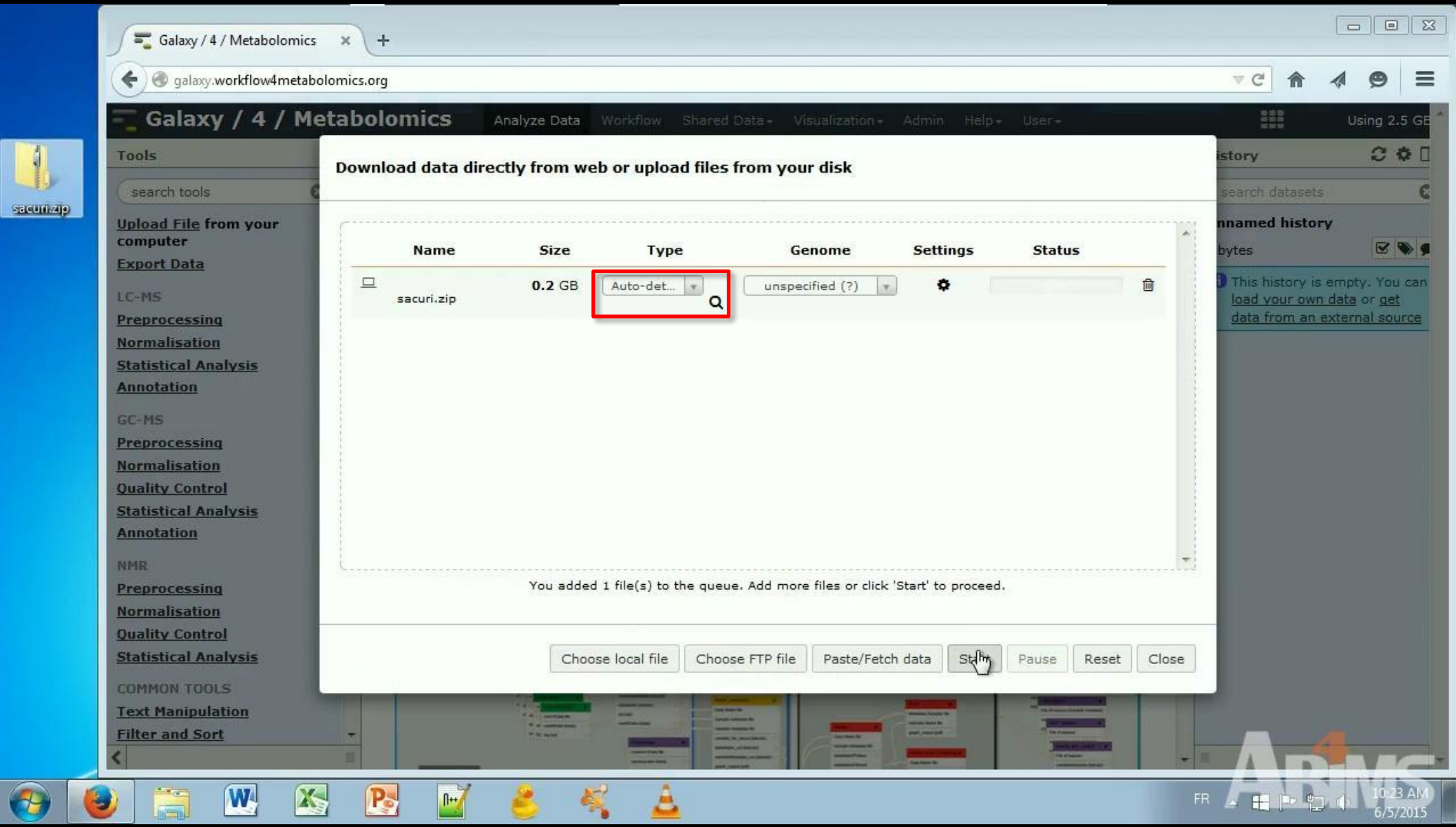
Data import

From local files



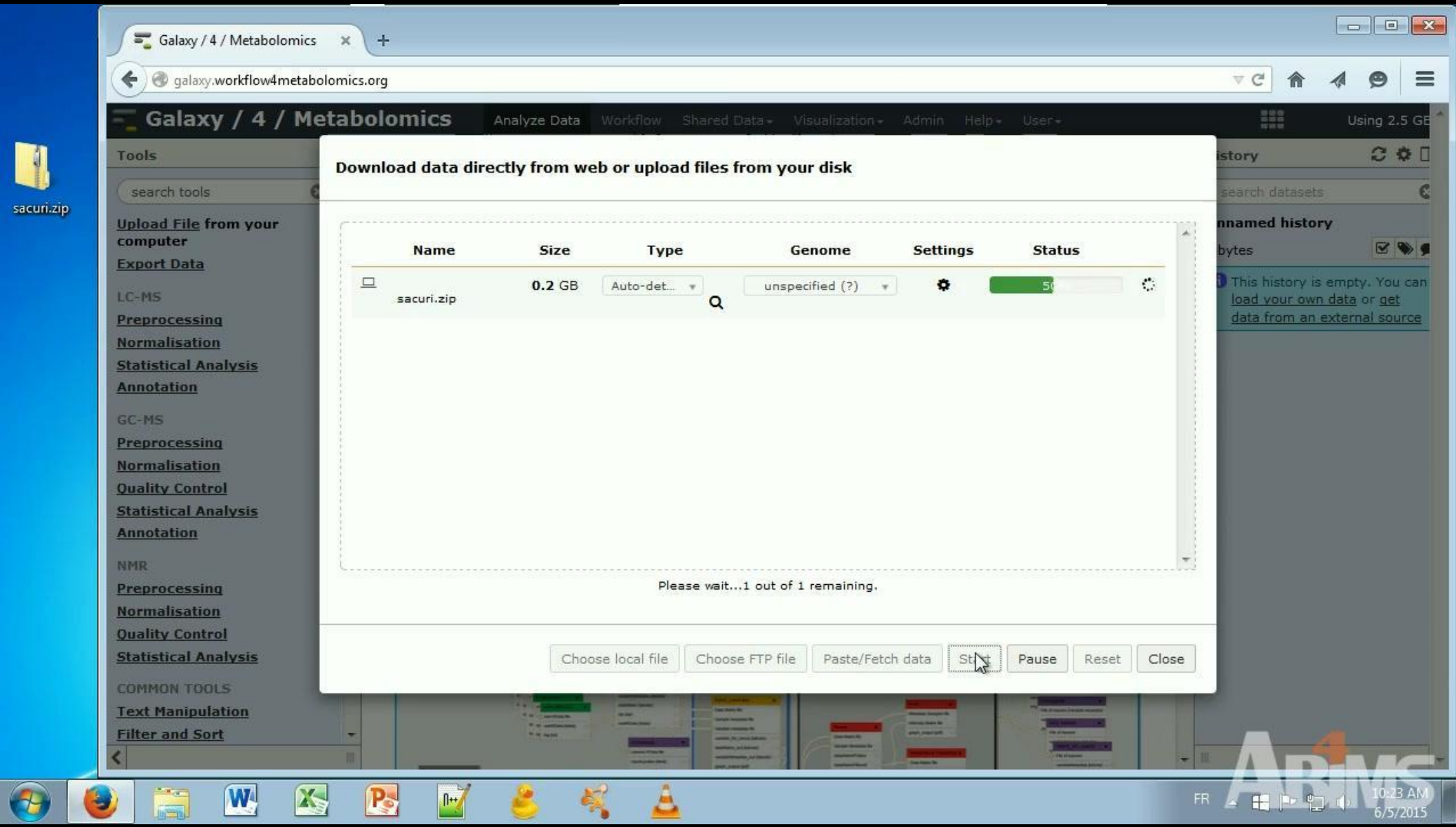
Data import

From local files



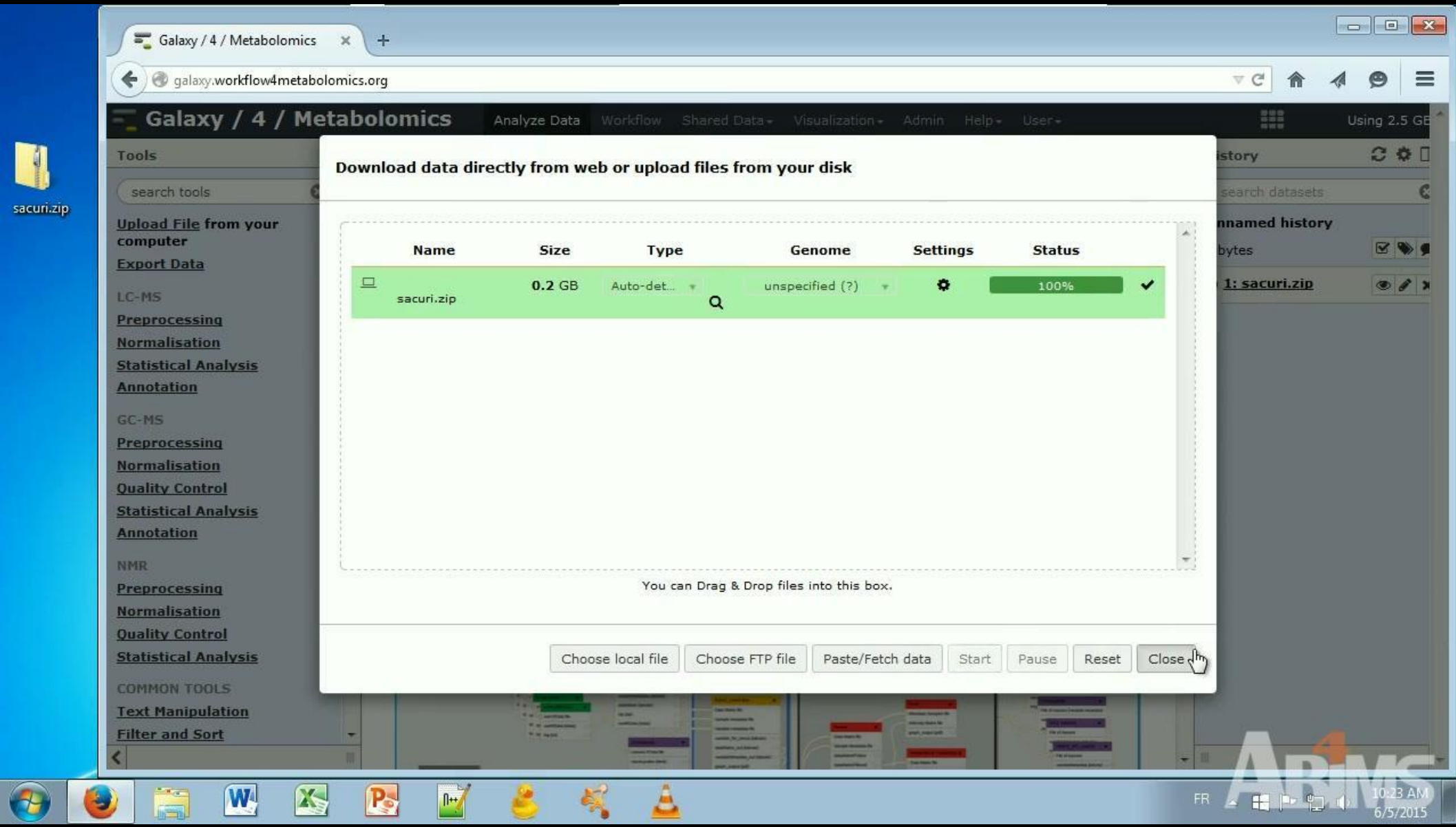
Data import

From local files



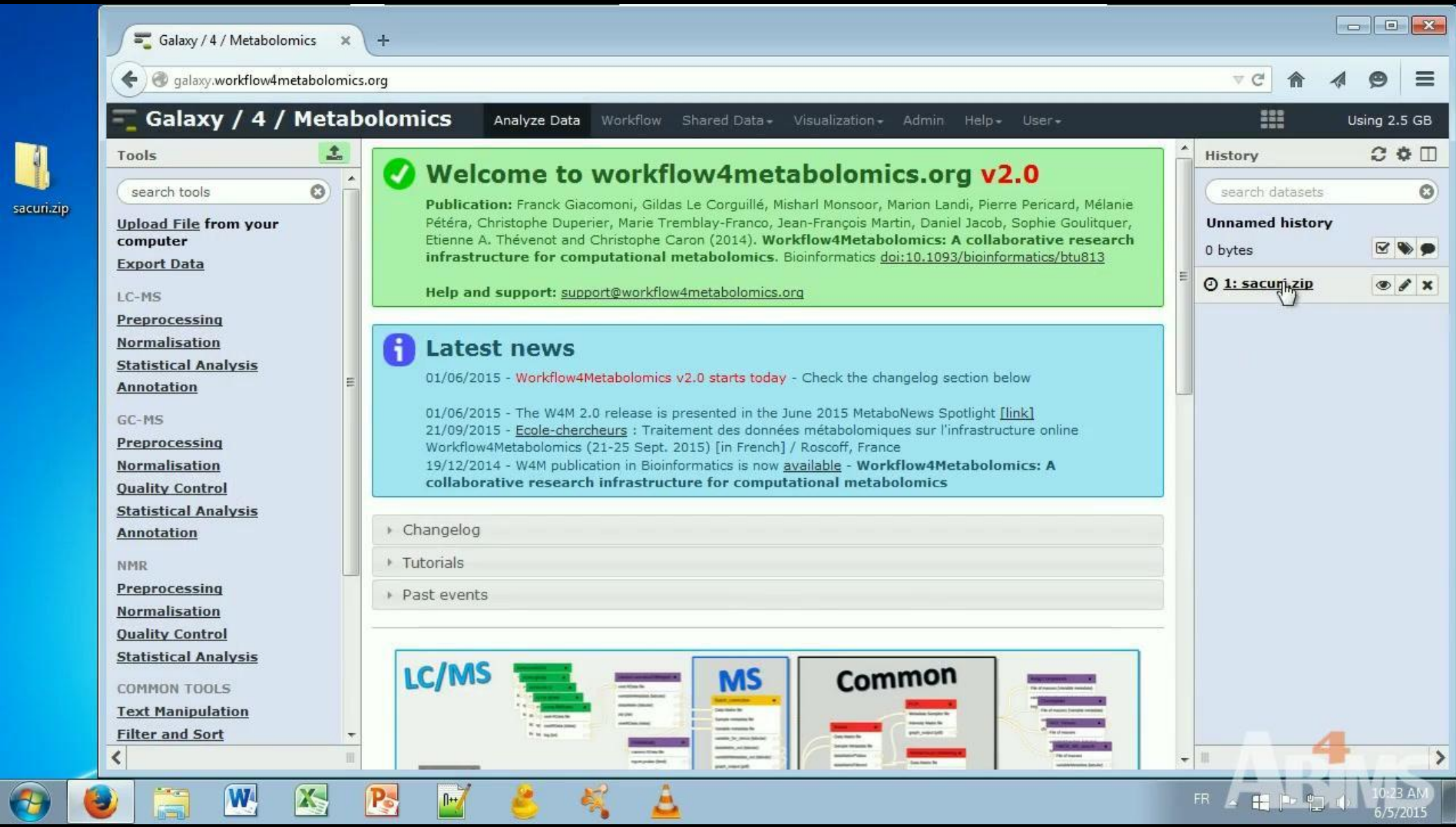
Data import

From local files



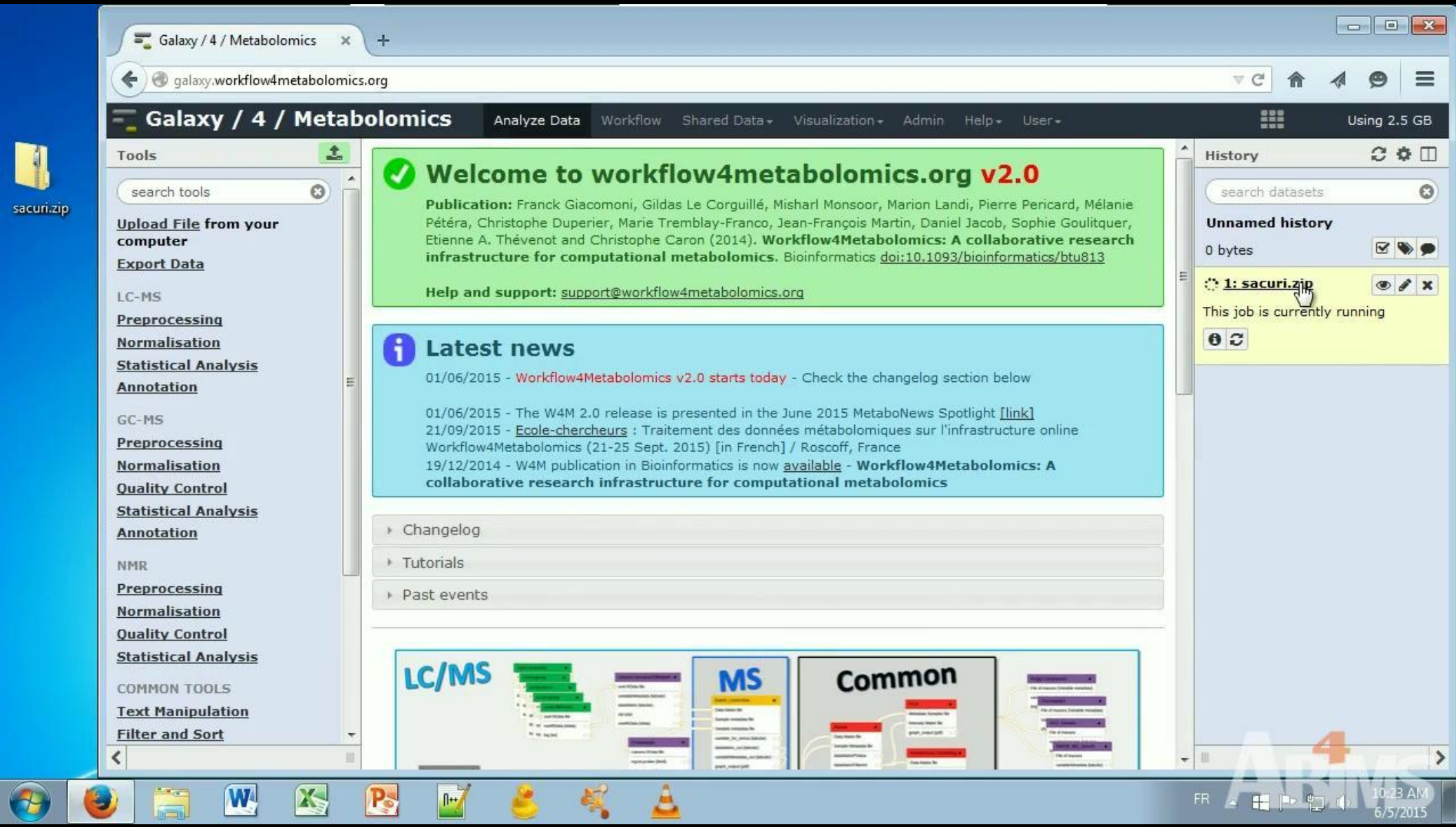
Data import

From local files



Data import

From local files

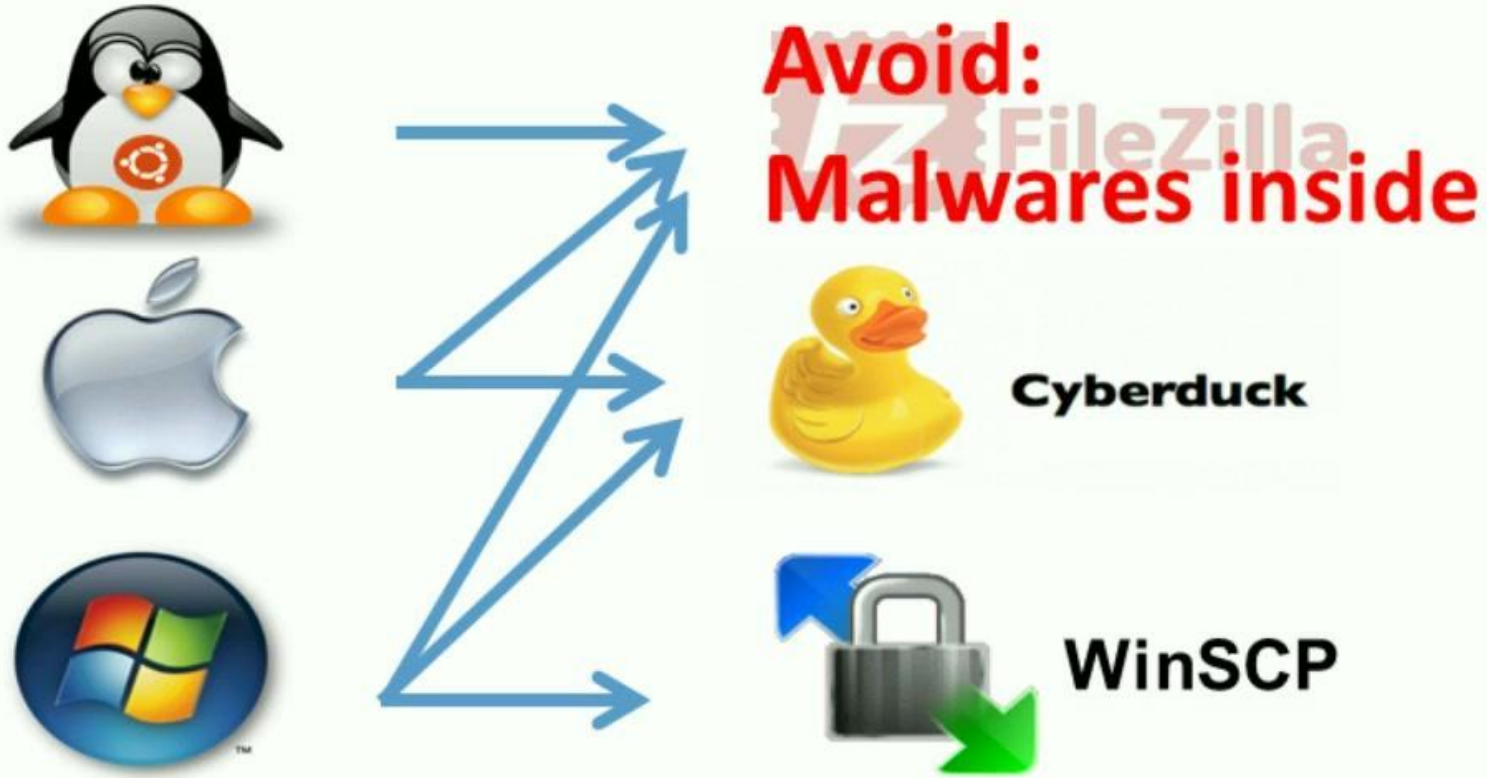


Step 1: Choose a FTP Client

DATA IMPORT USING FTP

STEP 1: CHOOSE A FTP CLIENT

STEP 1: CHOOSE A FTP CLIENT



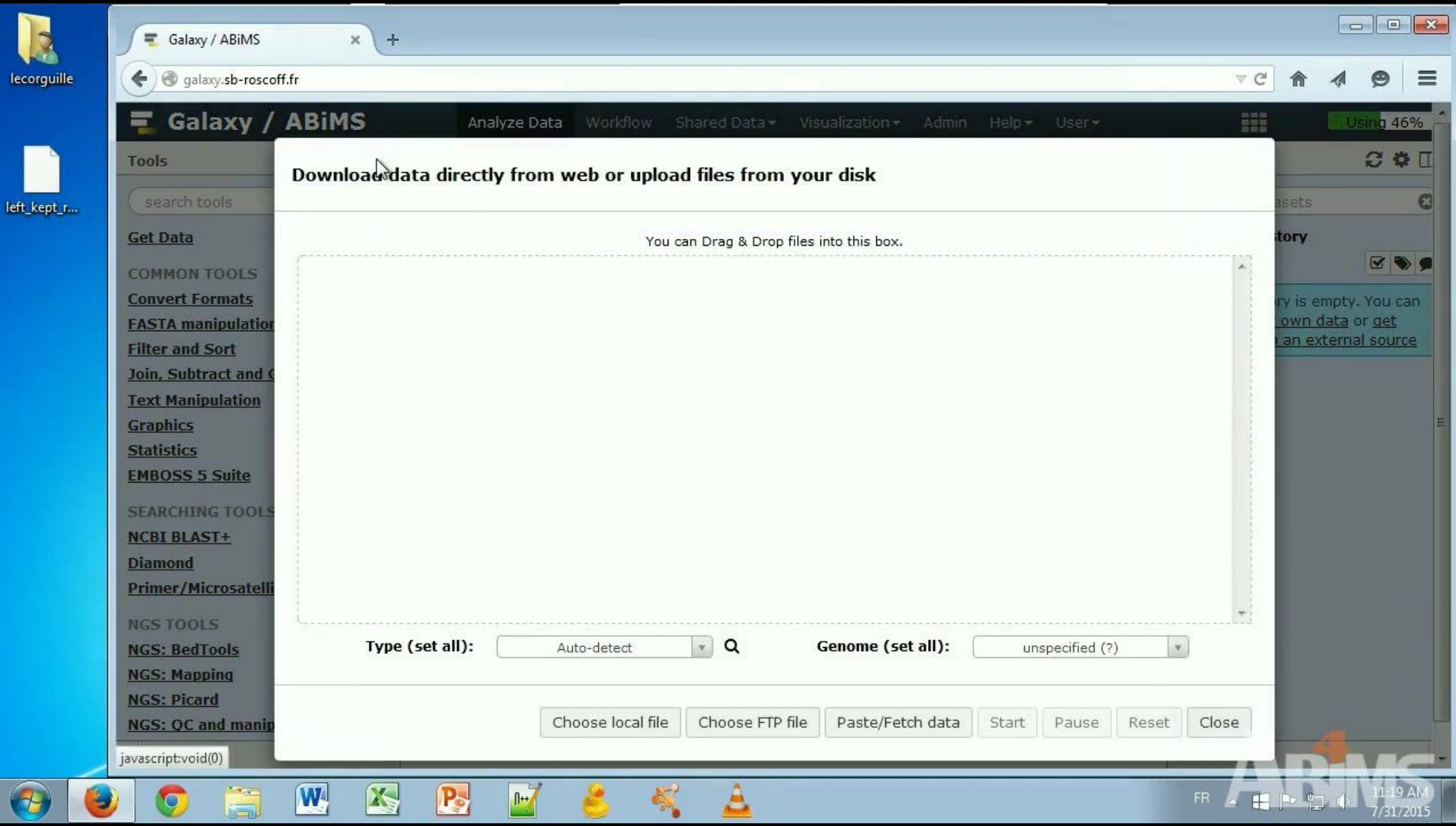
STEP 1: CHOOSE A FTP CLIENT

Step 2: Easy!

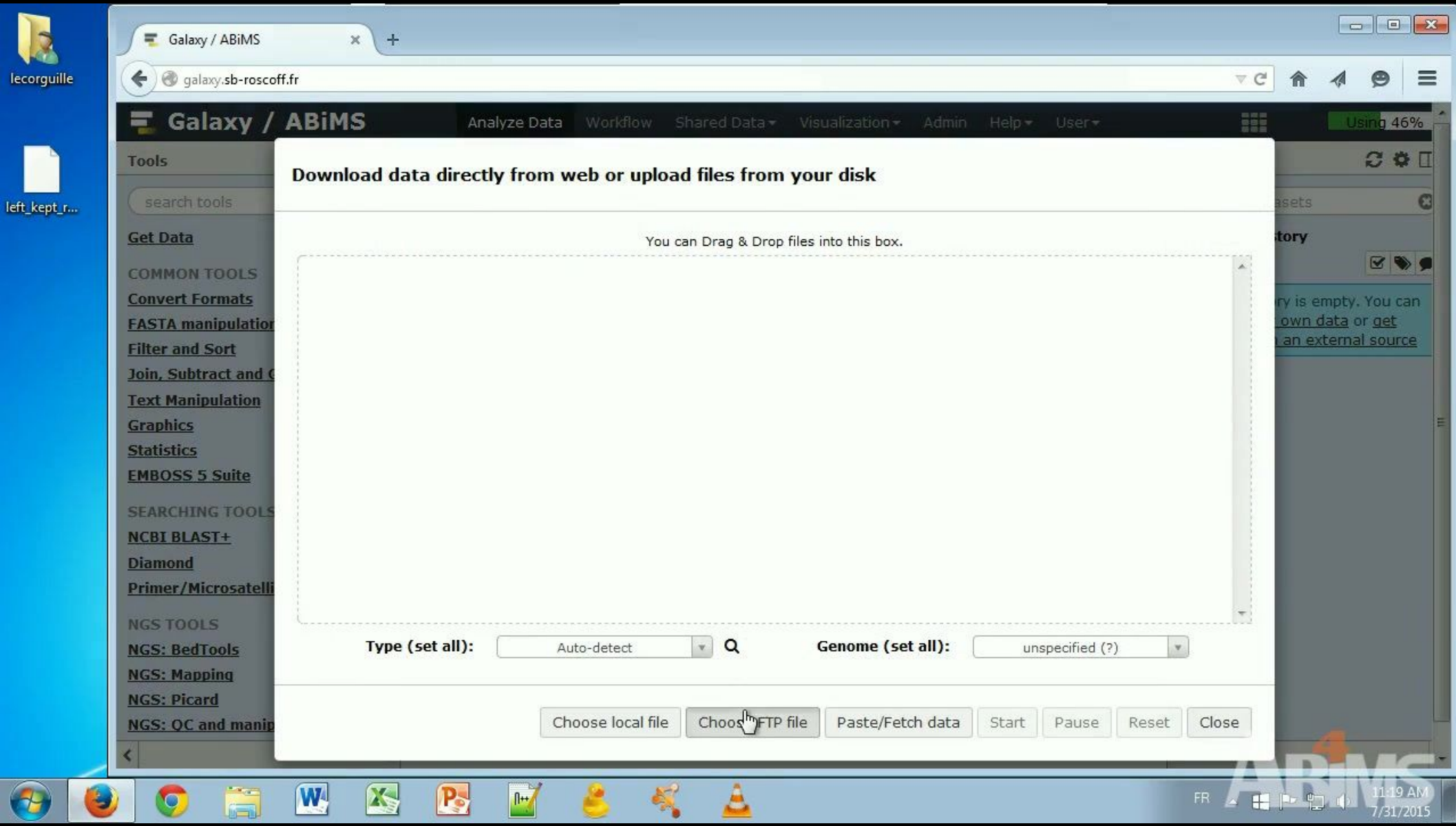
DATA IMPORT USING FTP

Data import using FTP

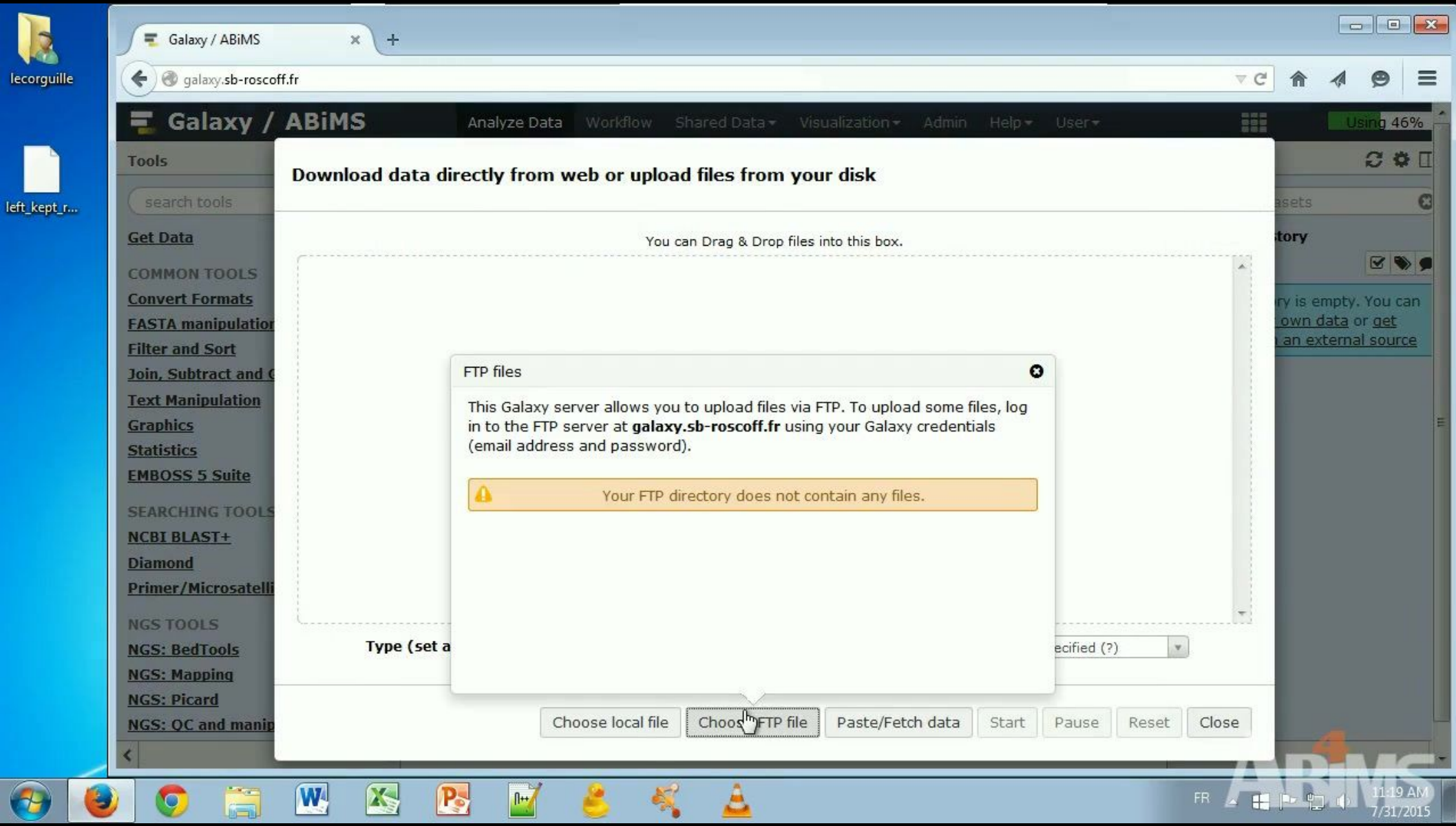
Data import using FTP



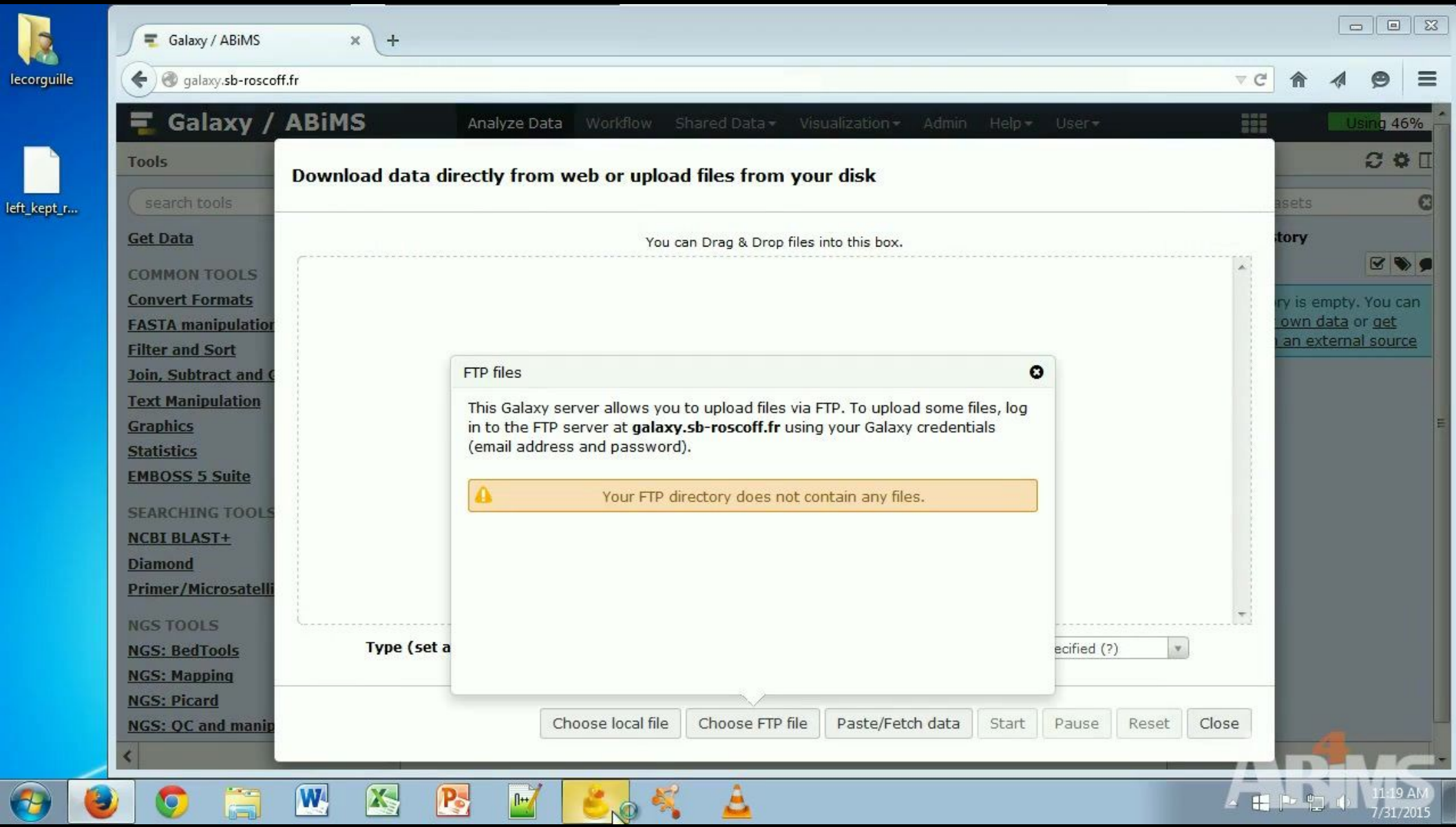
Data import using FTP



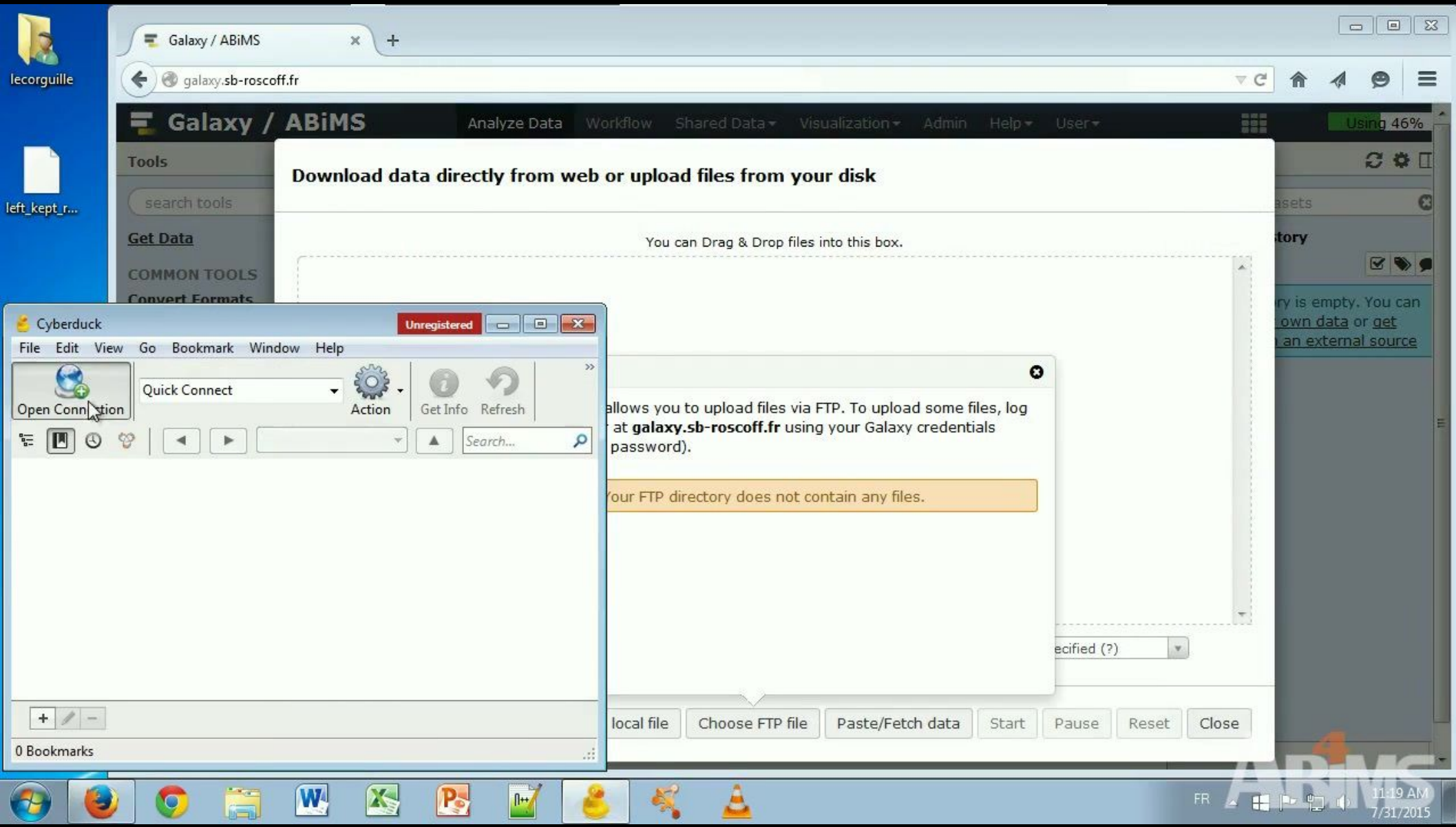
Data import using FTP



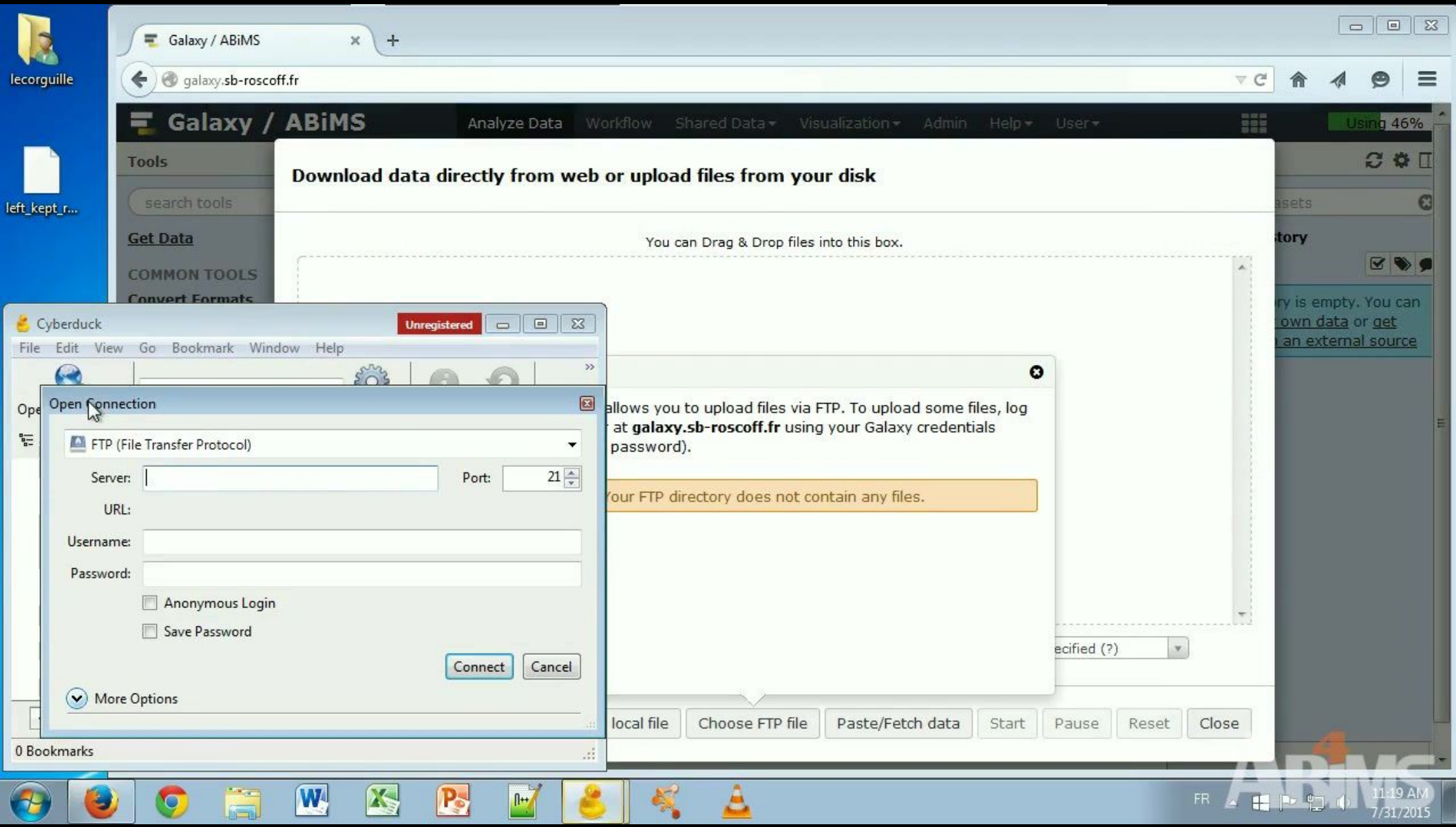
Data import using FTP



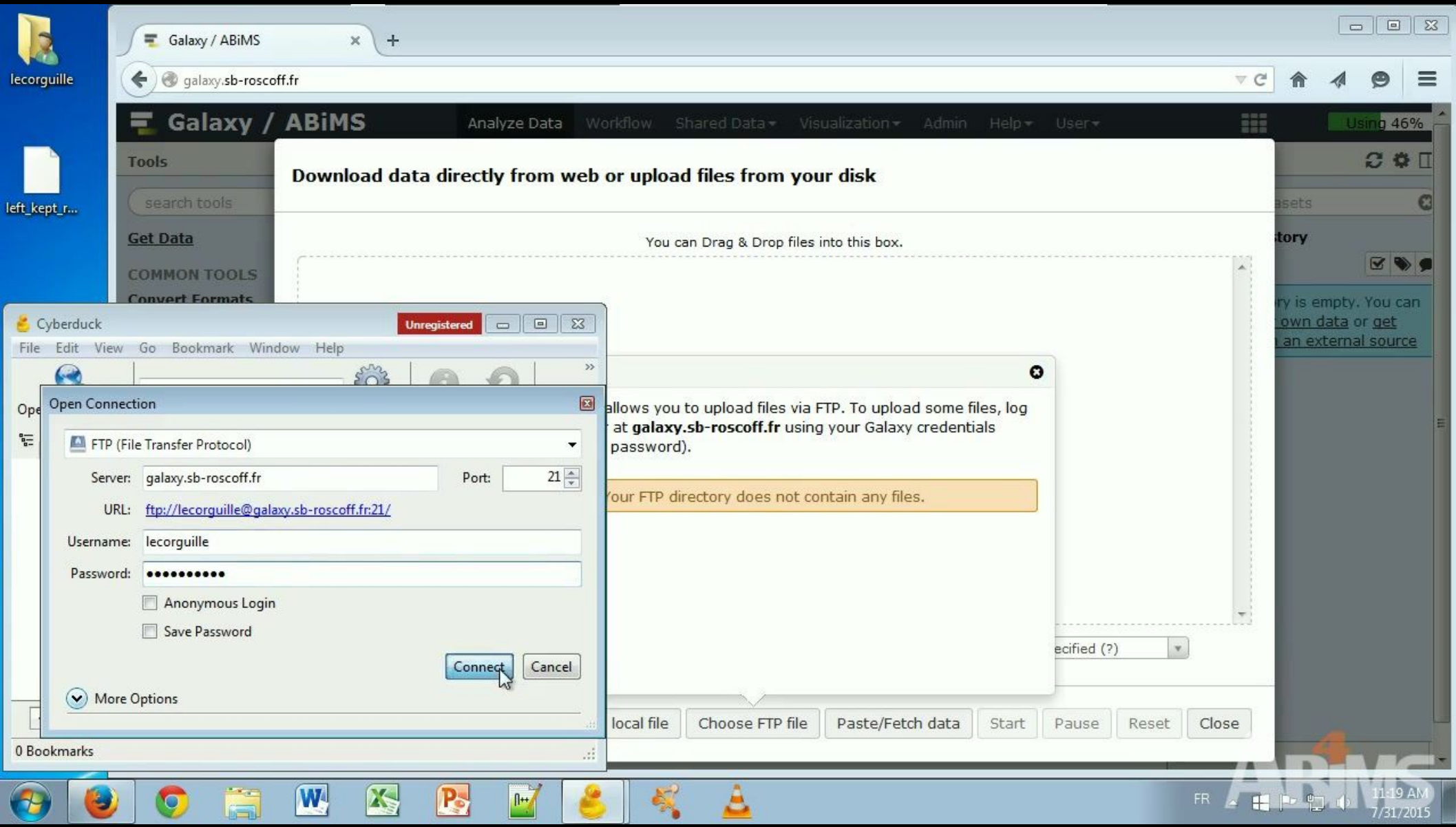
Data import using FTP



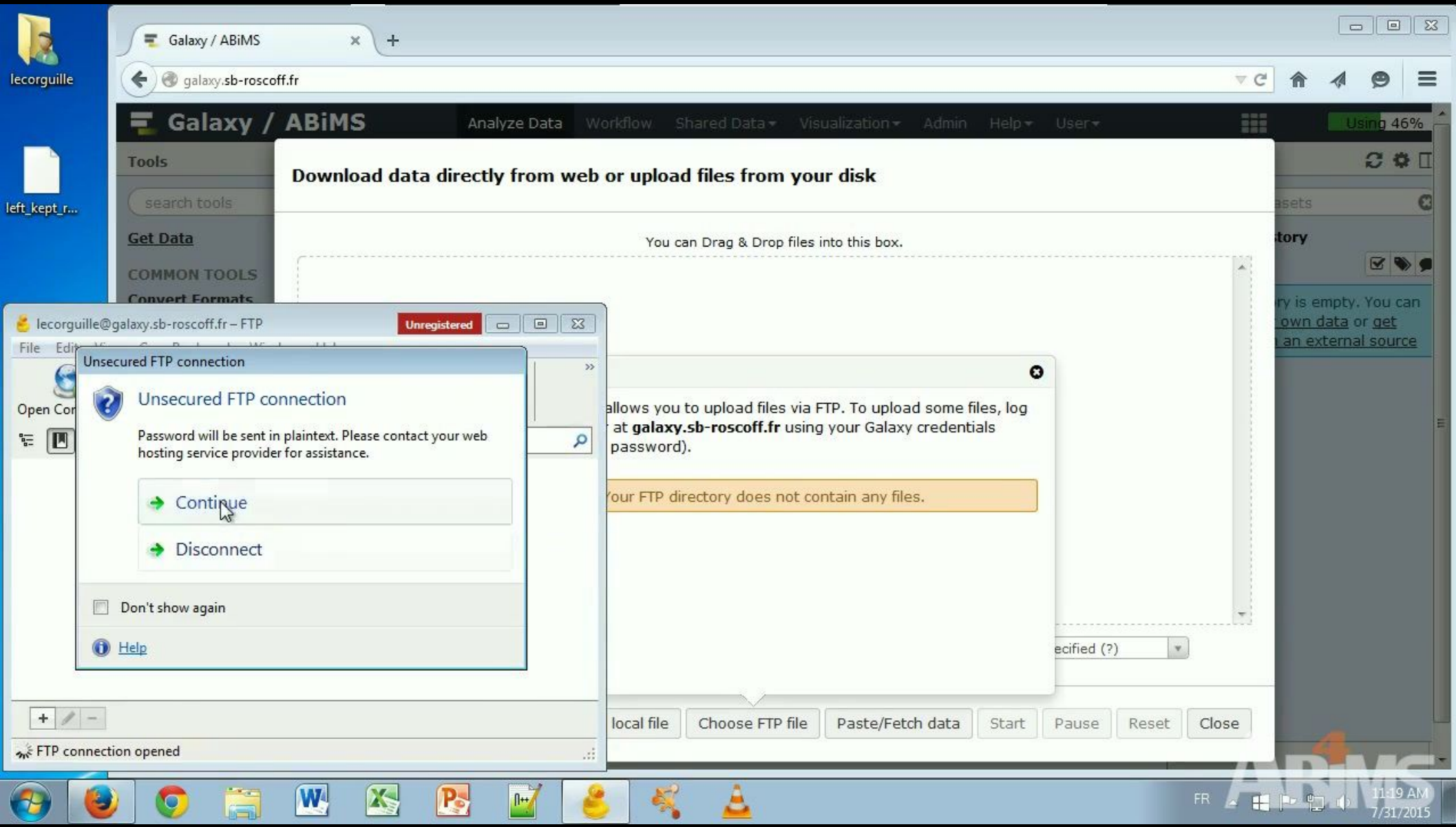
Data import using FTP



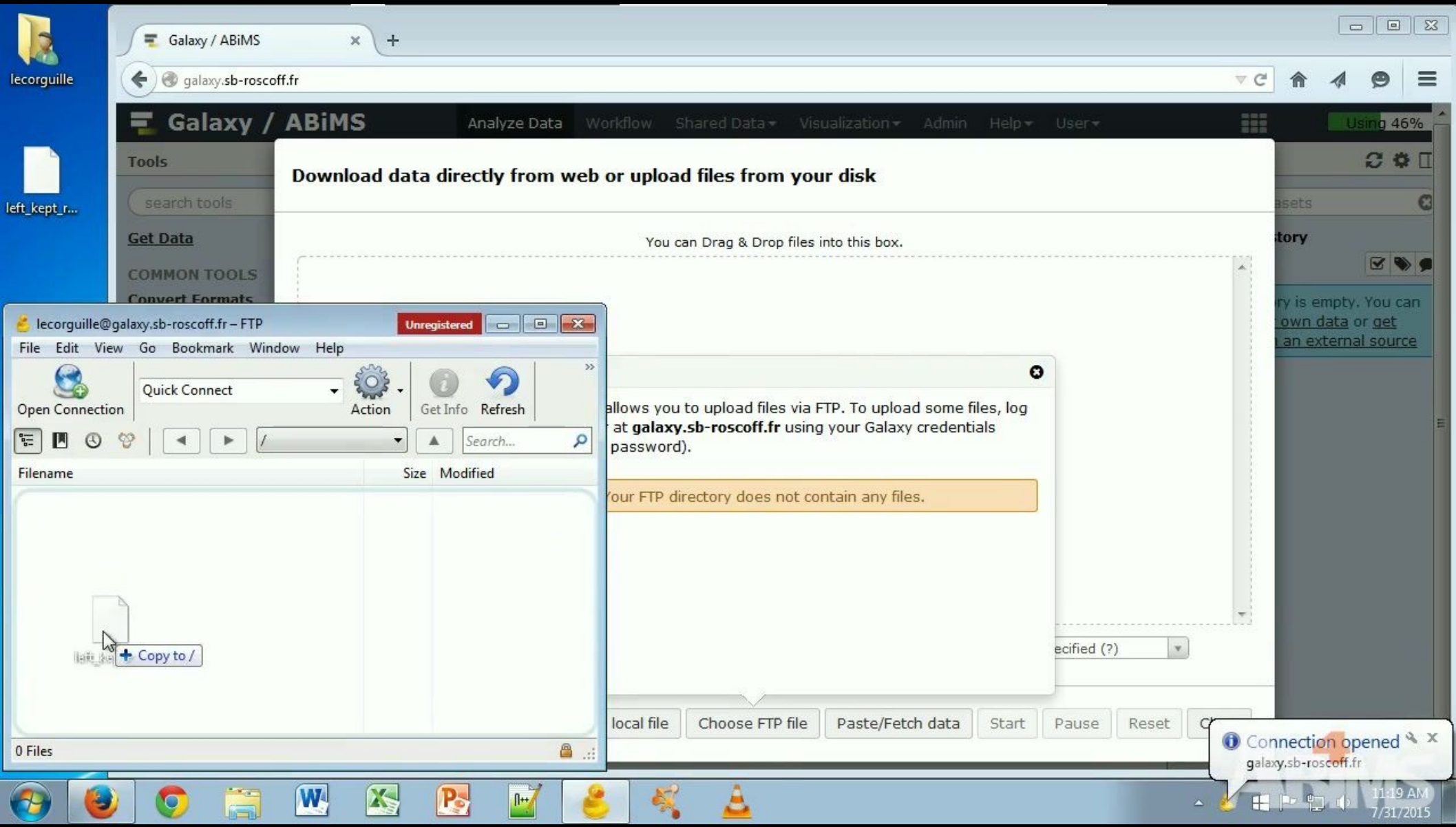
Data import using FTP



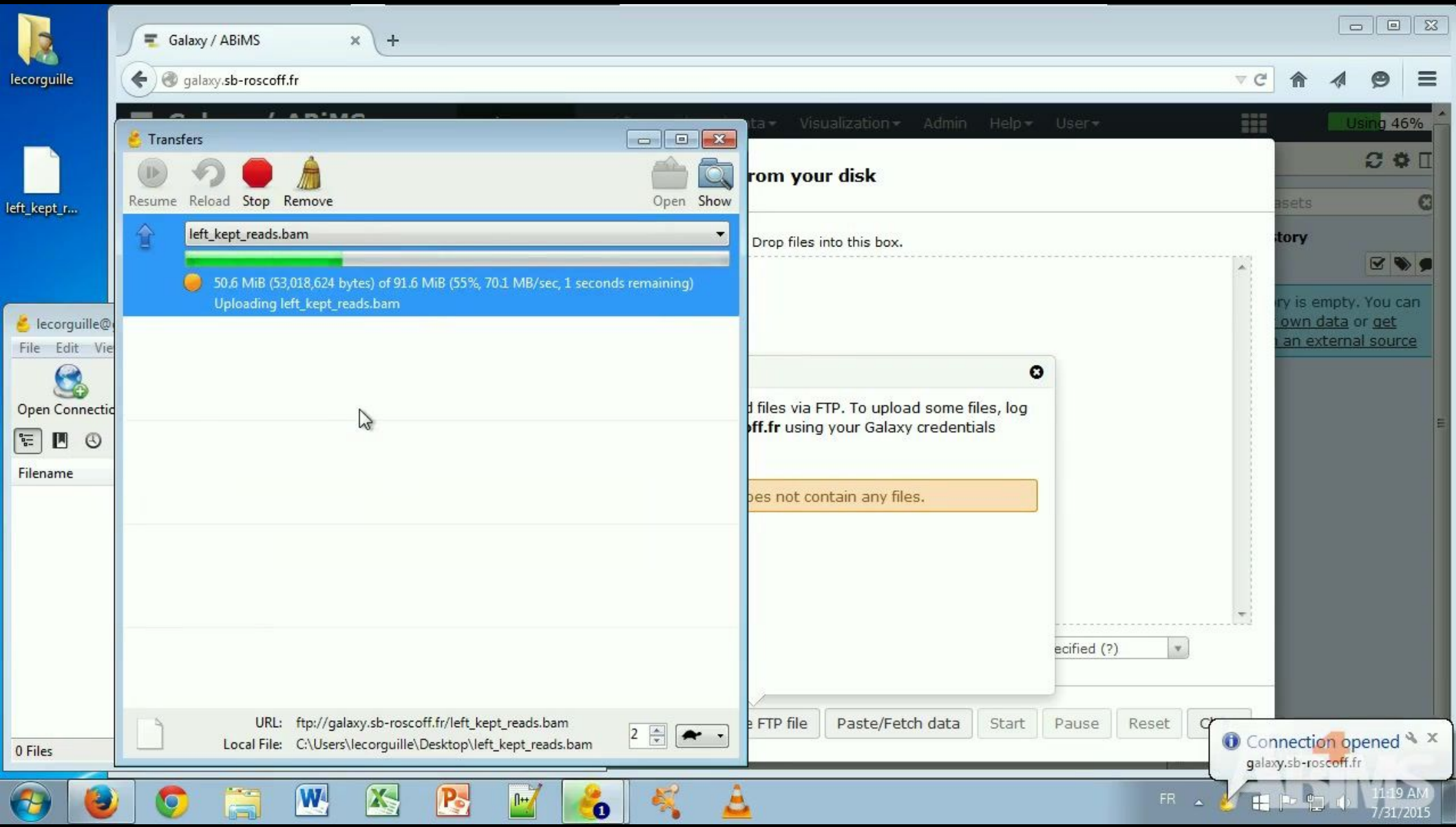
Data import using FTP



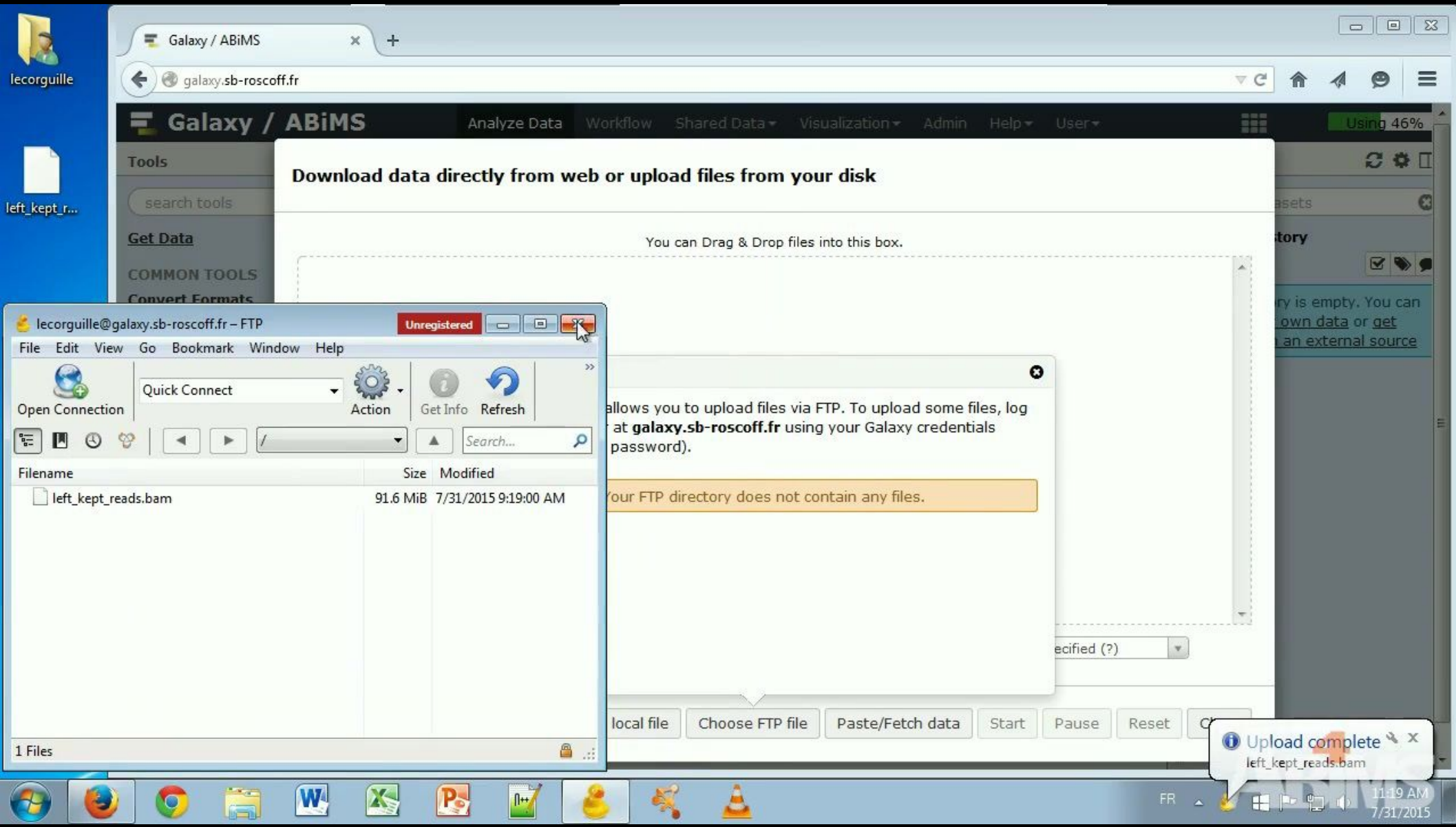
Data import using FTP



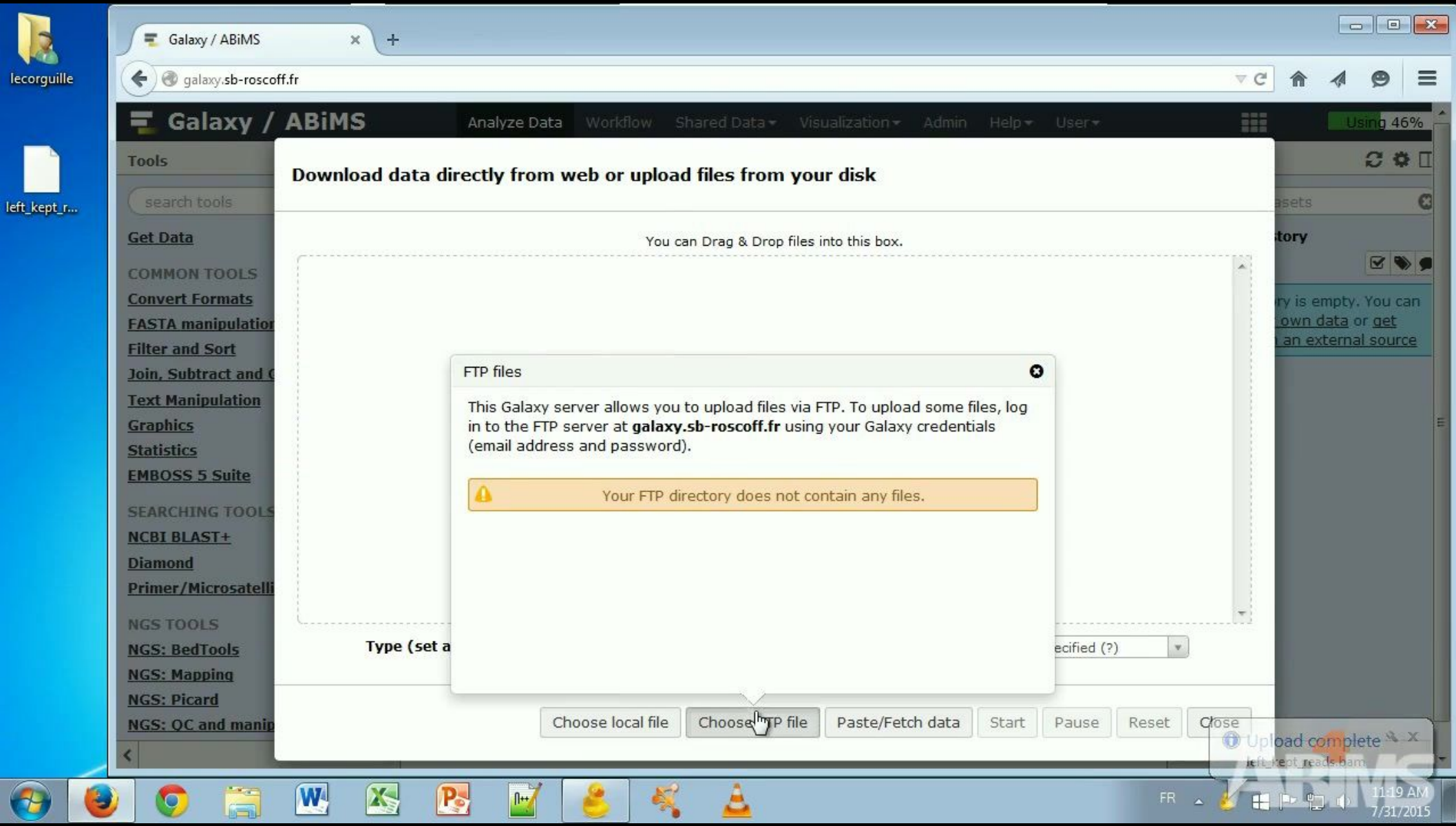
Data import using FTP



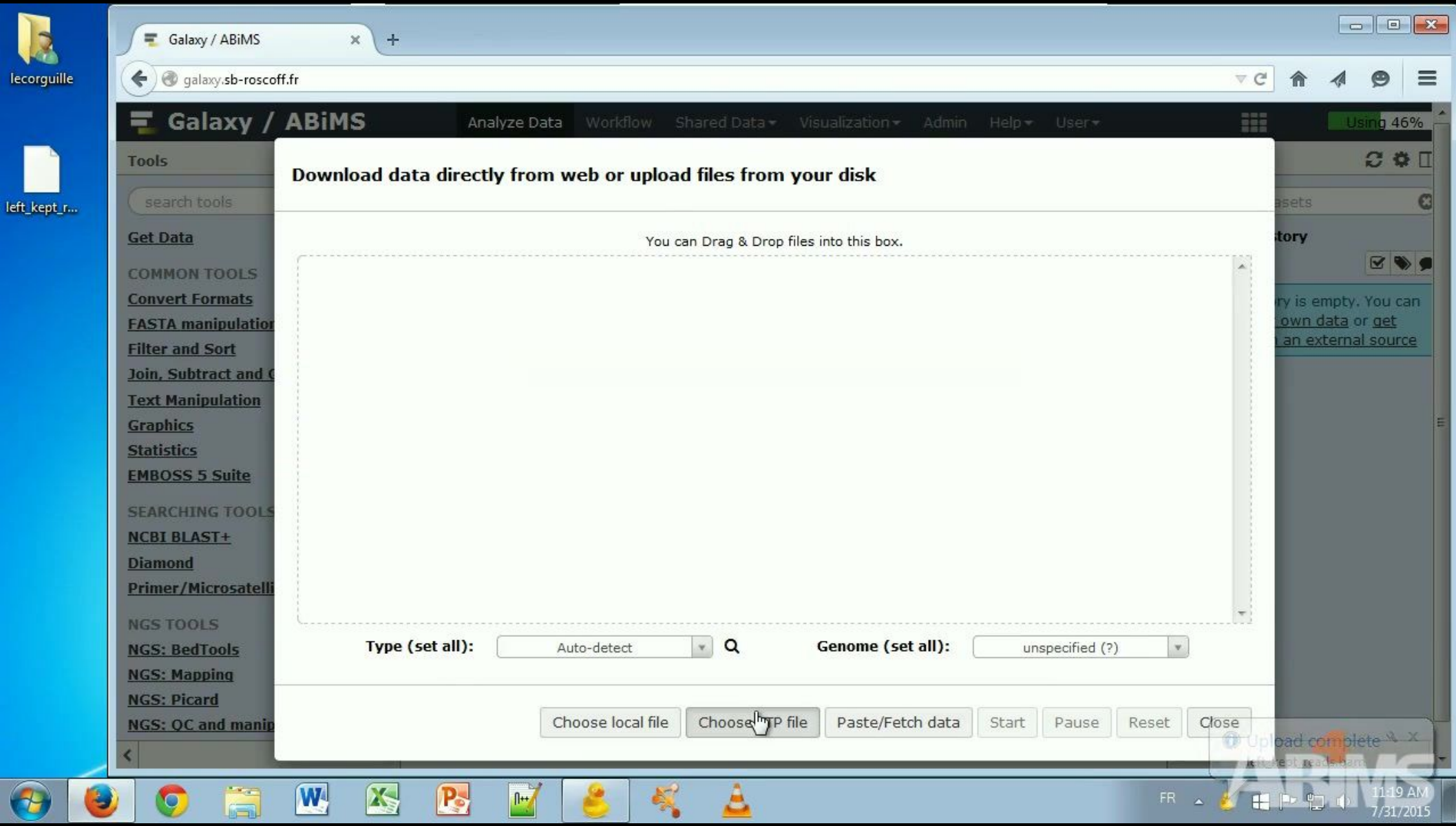
Data import using FTP



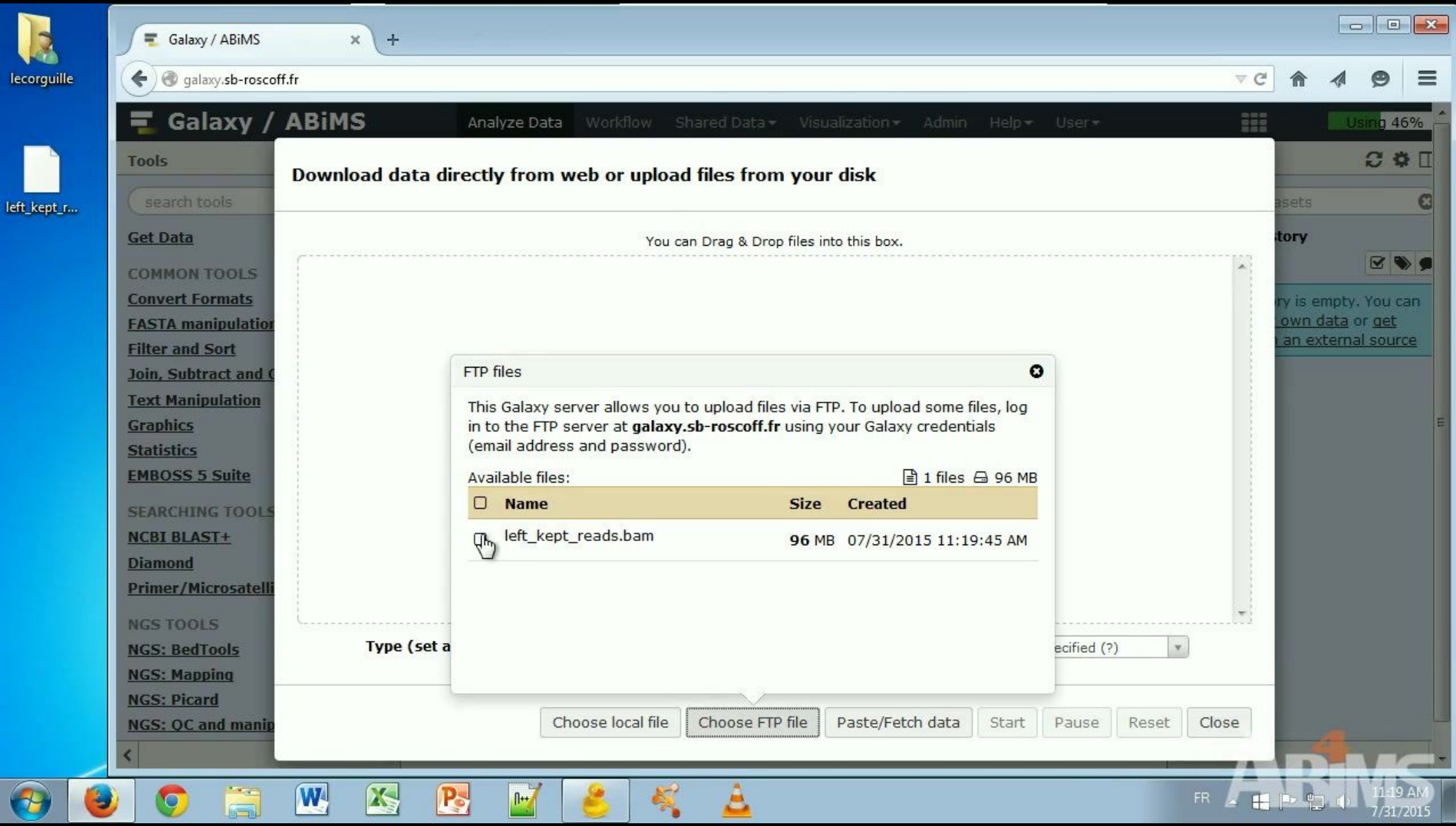
Data import using FTP



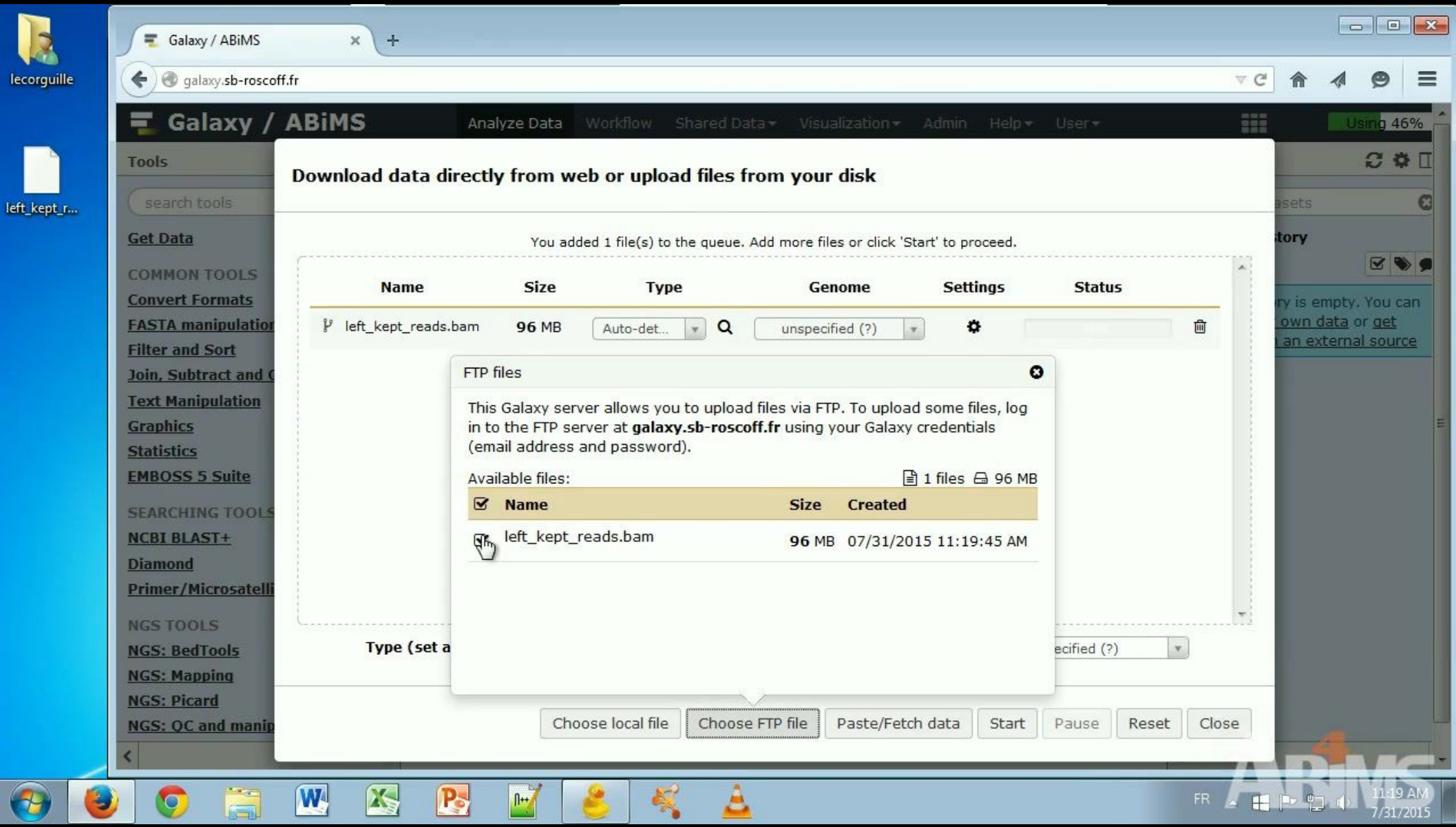
Data import using FTP



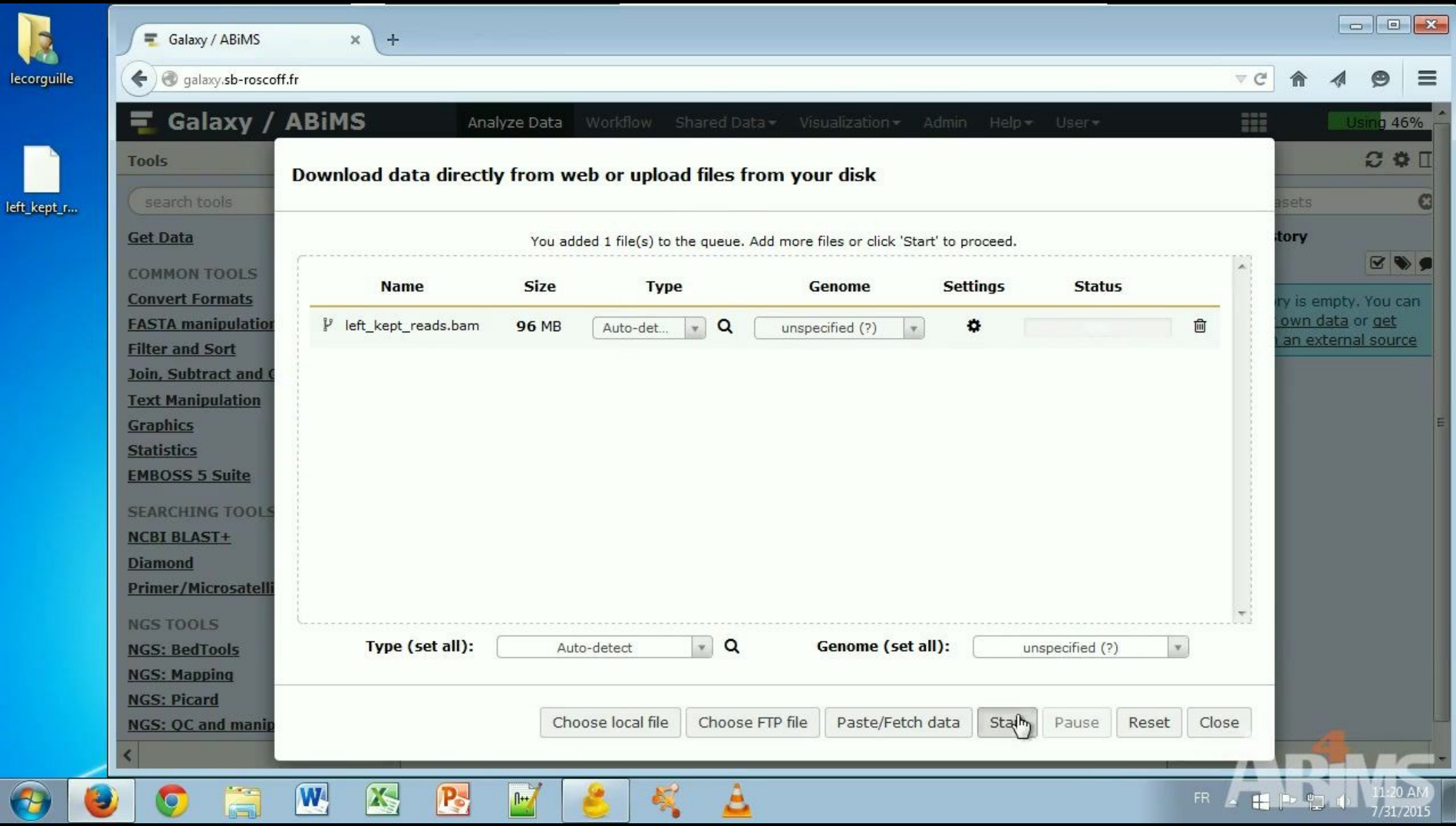
Data import using FTP



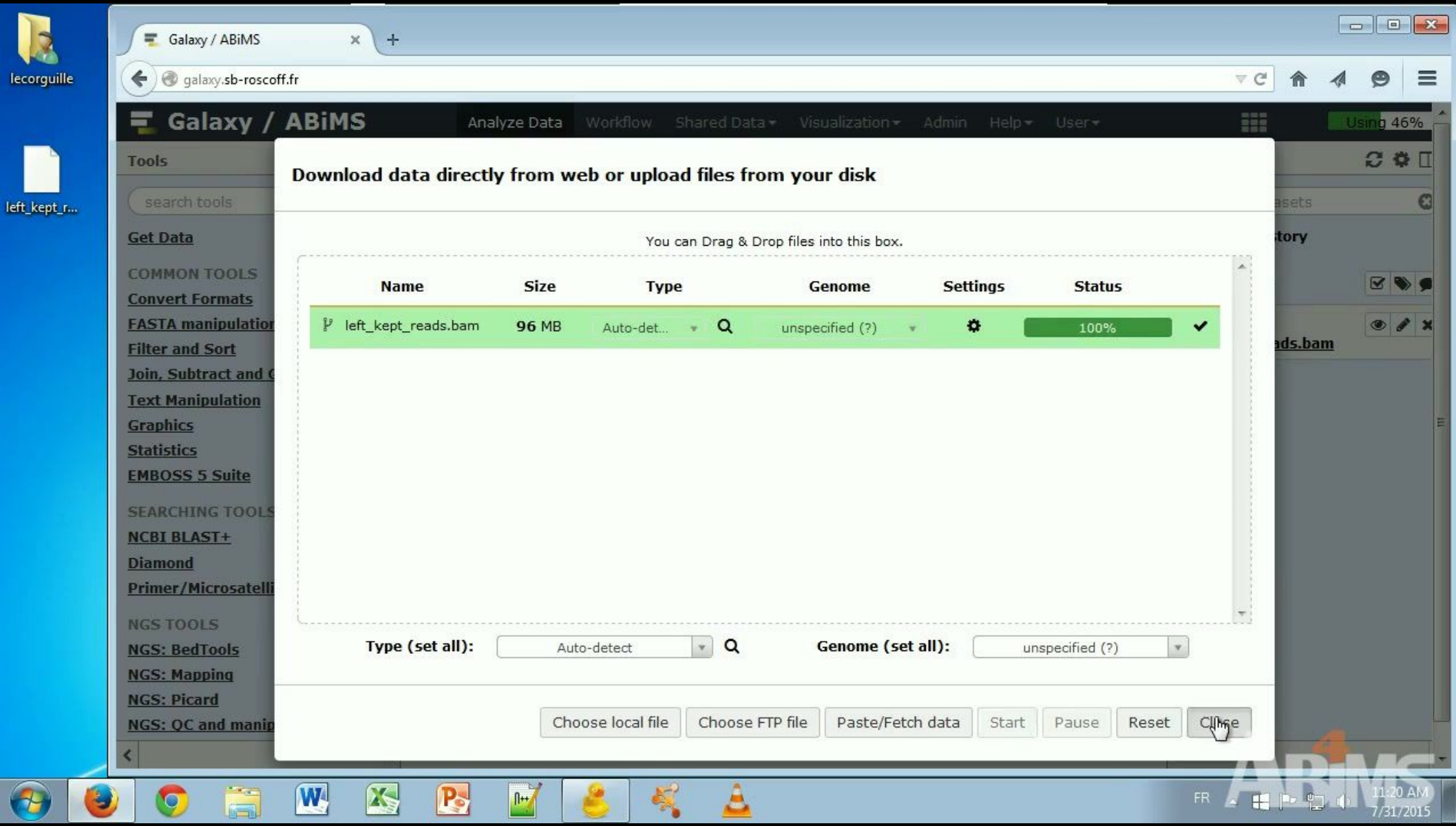
Data import using FTP



Data import using FTP



Data import using FTP



Data import using FTP

Data import using FTP

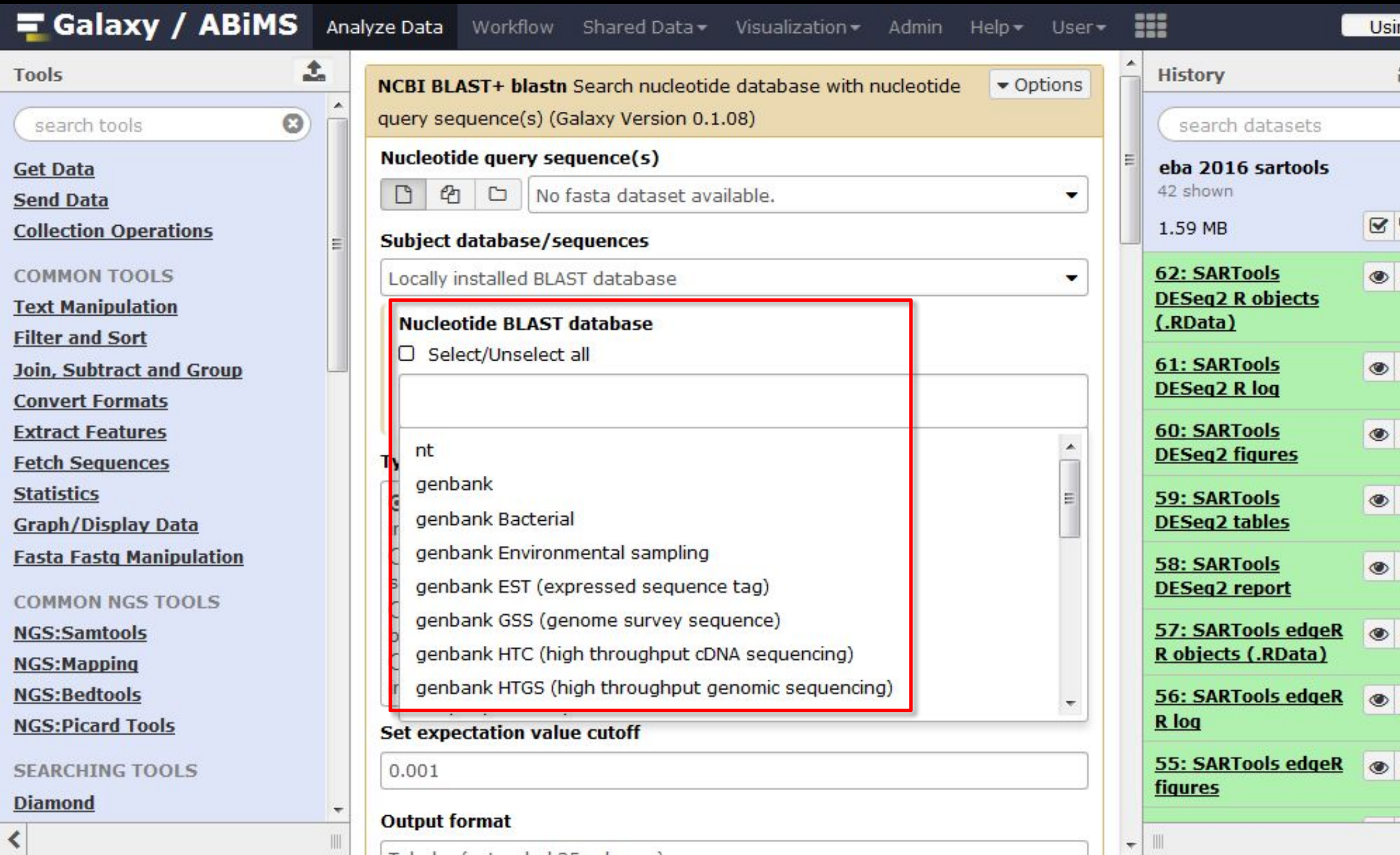
Data import using FTP

~~DATA IMPORT~~

~~Data import~~

For HUGE public resources: genome, databank …

--> Make a request to the support team



Hands-on

DATA IMPORT

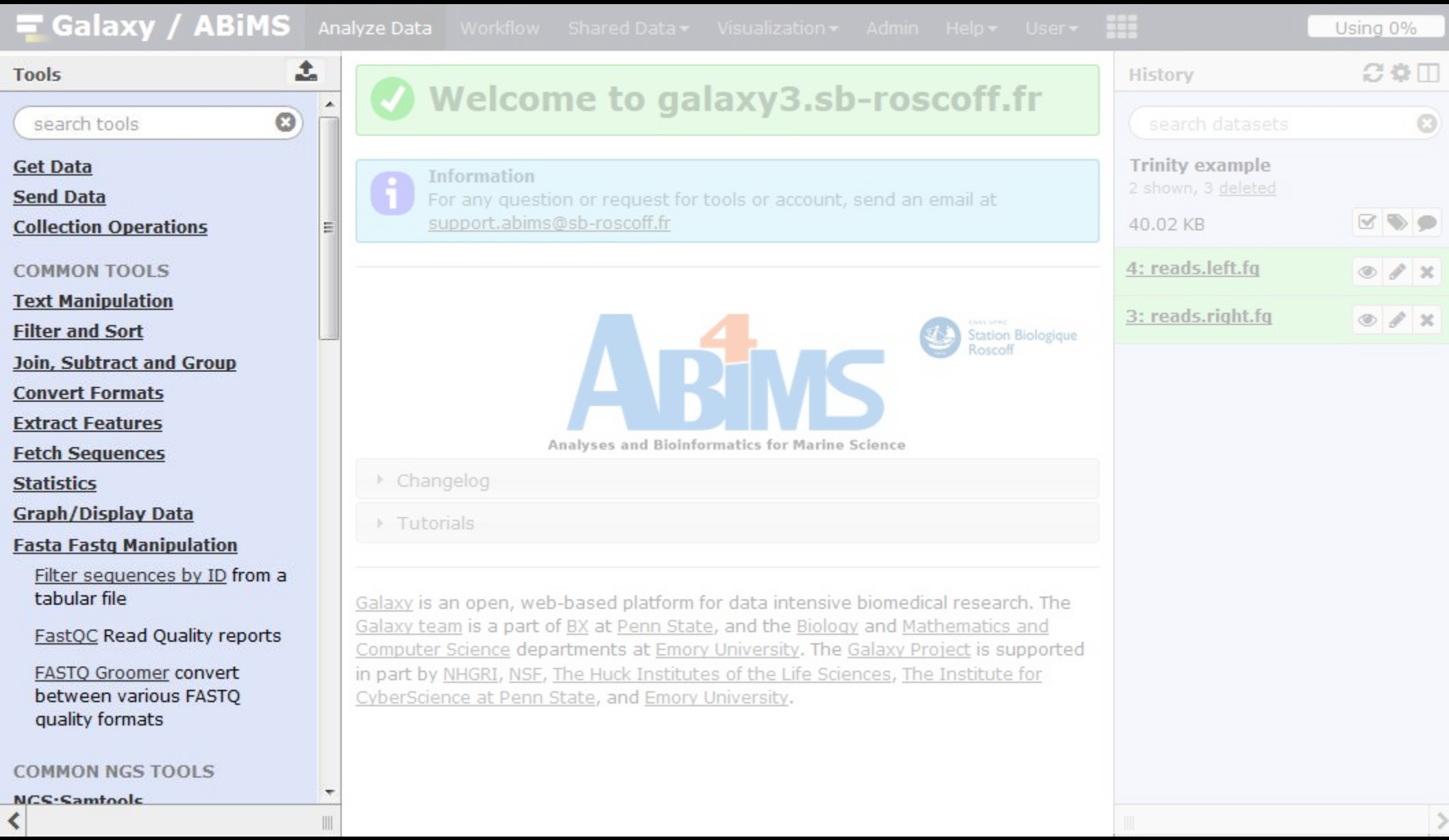
Data import – Hands-on

1. Fetch the files with your internet browser: https://tinyurl.com/galaxy-initiation-datasets

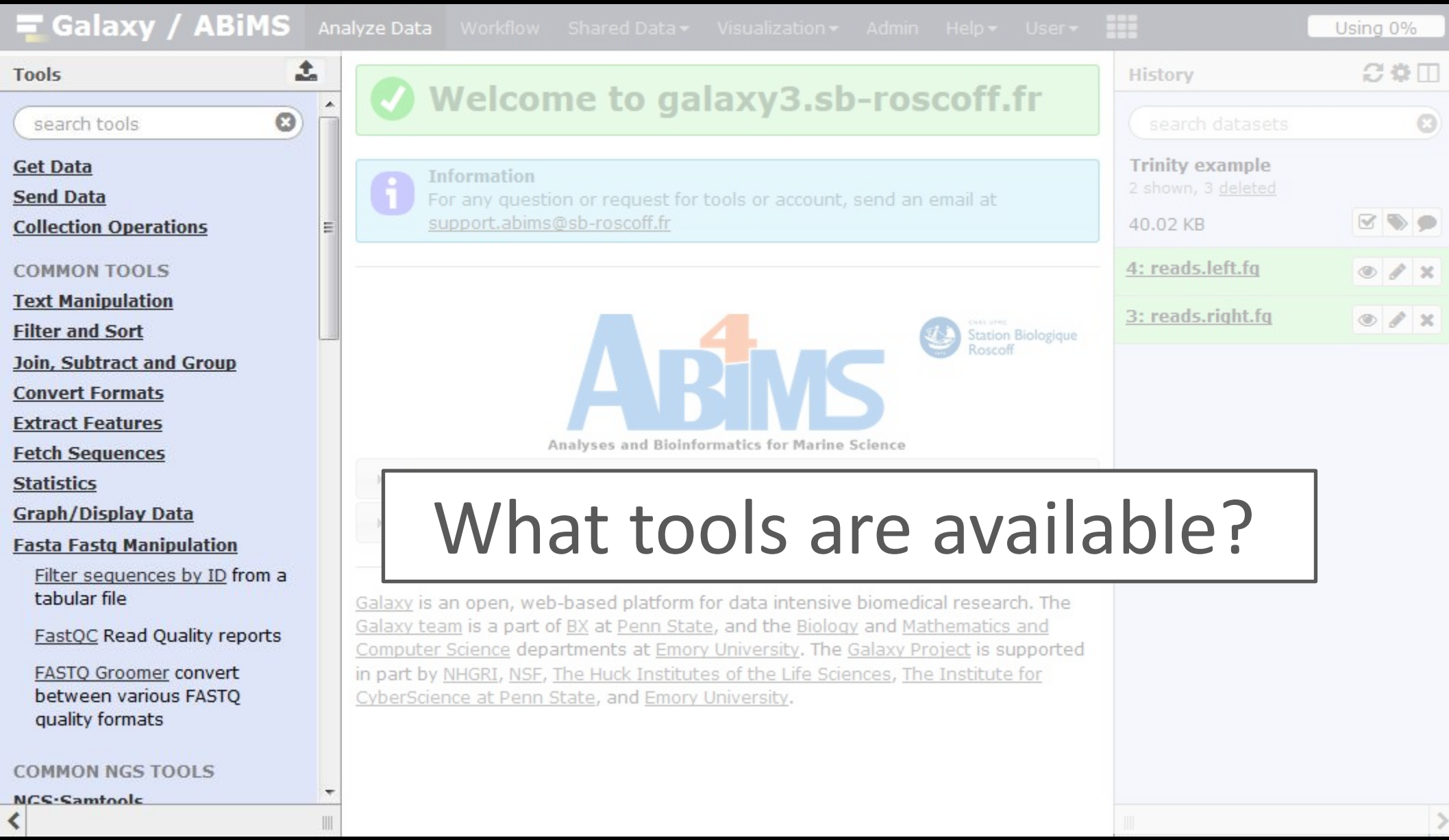
2. Upload files into Galaxy (exons.bed, snps.bed, repeats.bed)
   a. From disk
   b. Using FTP

TOOLS

Tools - panel


Galaxy / ABiMS
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 0%
Tools
search tools
Get Data
Send Data
Collection Operations
COMMON TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Statistics
Graph/Display Data
Fasta Fastq Manipulation
Filter sequences by ID from a tabular file
FastQC Read Quality reports
FASTQ Groomer convert between various FASTQ quality formats
COMMON NGS TOOLS
NGS:Samtools
Welcome to galaxy3.sb-roscoff.fr
Information
For any question or request for tools or account, send an email at support.abims@sb-roscoff.fr
Analyses and Bioinformatics for Marine Science
Station Biologique Roscoff
Changelog
Tutorials
Galaxy is an open, web-based platform for data intensive biomedical research. 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 NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.
History
search datasets
Trinity example
2 shown, 3 deleted
40.02 KB
4: reads.left.fq
3: reads.right.fq

Tools - panel


Galaxy / ABiMS
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 0%
Tools
search tools
Get Data
Send Data
Collection Operations
COMMON TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Statistics
Graph/Display Data
Fasta Fastq Manipulation
Filter sequences by ID from a tabular file
FastQC Read Quality reports
FASTQ Groomer convert between various FASTQ quality formats
COMMON NGS TOOLS
NGS:Samtools
Welcome to galaxy3.sb-roscoff.fr
Information
For any question or request for tools or account, send an email at support.abims@sb-roscoff.fr
Analyses and Bioinformatics for Marine Science
Station Biologique Roscoff
What tools are available?
Galaxy is an open, web-based platform for data intensive biomedical research. 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 NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.
History
search datasets
Trinity example
2 shown, 3 deleted
40.02 KB
4: reads.left.fq
3: reads.right.fq

Tools - panel

125+ platforms for using Galaxy: https://galaxyproject.org/use

[Servers, clouds, deployable resources]

Tools - panel

125+ platforms for using Galaxy: https://galaxyproject.org/use
[Servers, clouds, deployable resources]

GENERAL PURPOSE:

usegalaxy.org, usegalaxy.eu, usegalaxy.fr, usegalaxy.org.au

Tools - panel

125+ platforms for using Galaxy: https://galaxyproject.org/use
[Servers, clouds, deployable resources]

GENERAL PURPOSE:

usegalaxy.org, usegalaxy.eu, usegalaxy.fr, usegalaxy.org.au

DOMAIN SPECIFIC:

ABiMS

RNAseq: http://galaxy.sb-roscoff.fr
SBR tools: http://webtools.sb-roscoff.fr
Metagenomics: http://galaxy4metab.sb-roscoff.fr

Metabolomics: 4 Wm Workflow4metabolomics

Proteomics: ProteoRE

ChIP-seq: GalaxEast

nebula ouris.fr

Tools - panel

125+ platforms for using Galaxy: https://galaxyproject.org/use
[Servers, clouds, deployable resources]

GENERAL PURPOSE:

usegalaxy.org, usegalaxy.eu, usegalaxy.fr, usegalaxy.org.au

DOMAIN SPECIFIC:

RNAseq: http://galaxy.sb-roscoff.fr
SBR tools: http://webtools.sb-roscoff.fr
Metagenomics: http://galaxy4metab.sb-roscoff.fr

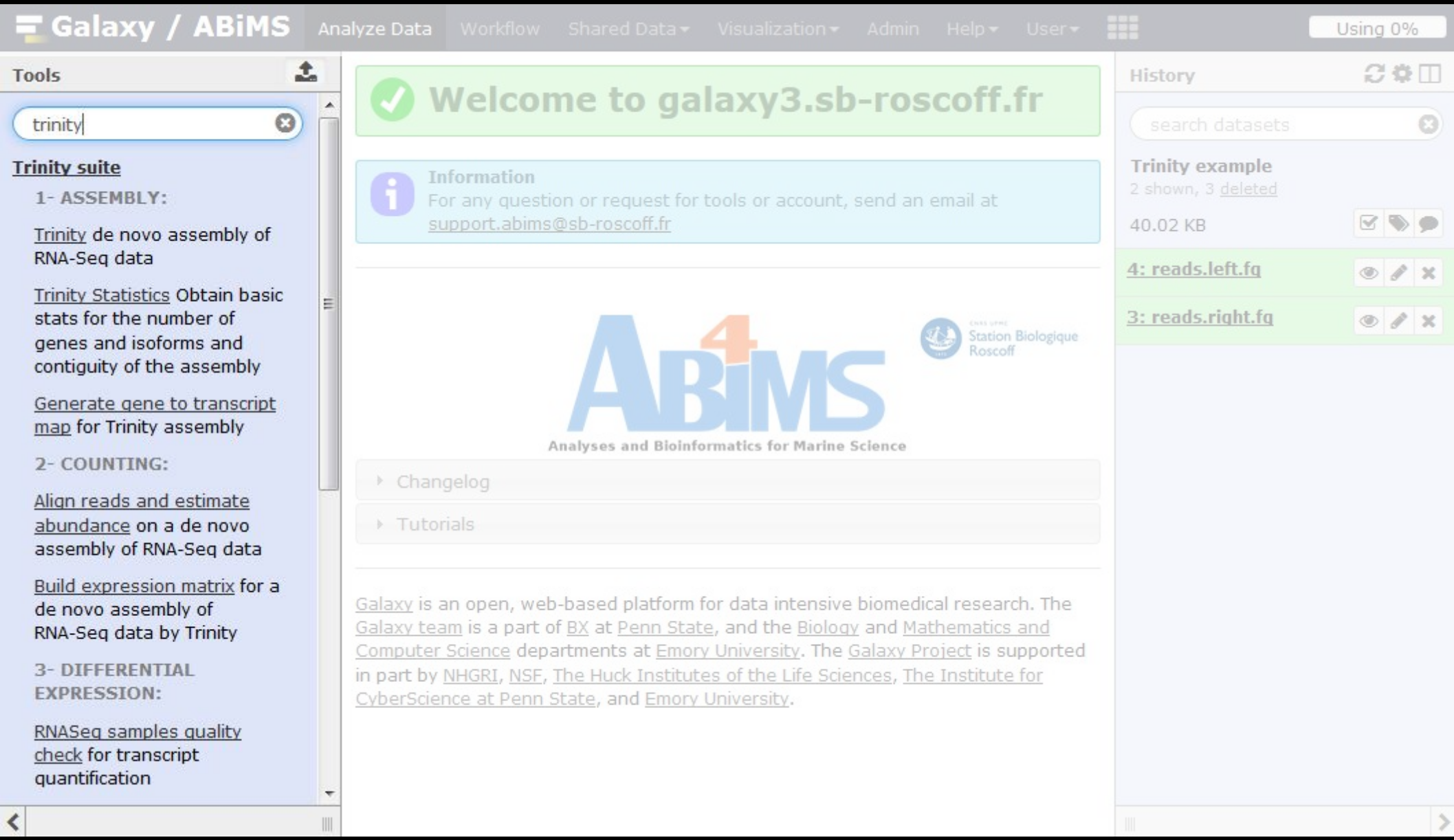
Metabolomics:

Proteomics:

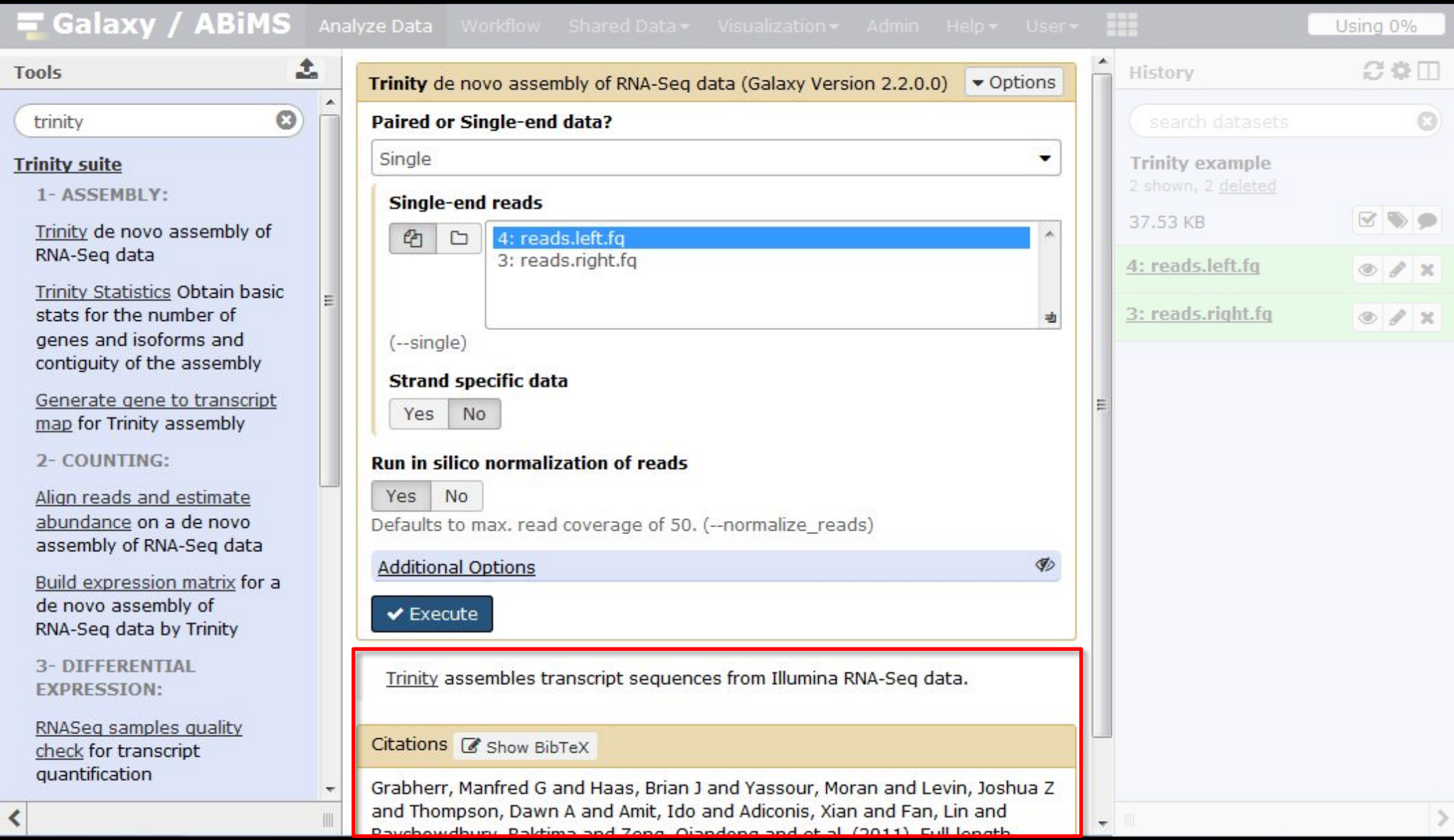
ChIP-seq:

Catalog of French Galaxy tools:
http://galaxycat.france-bioinformatique.fr

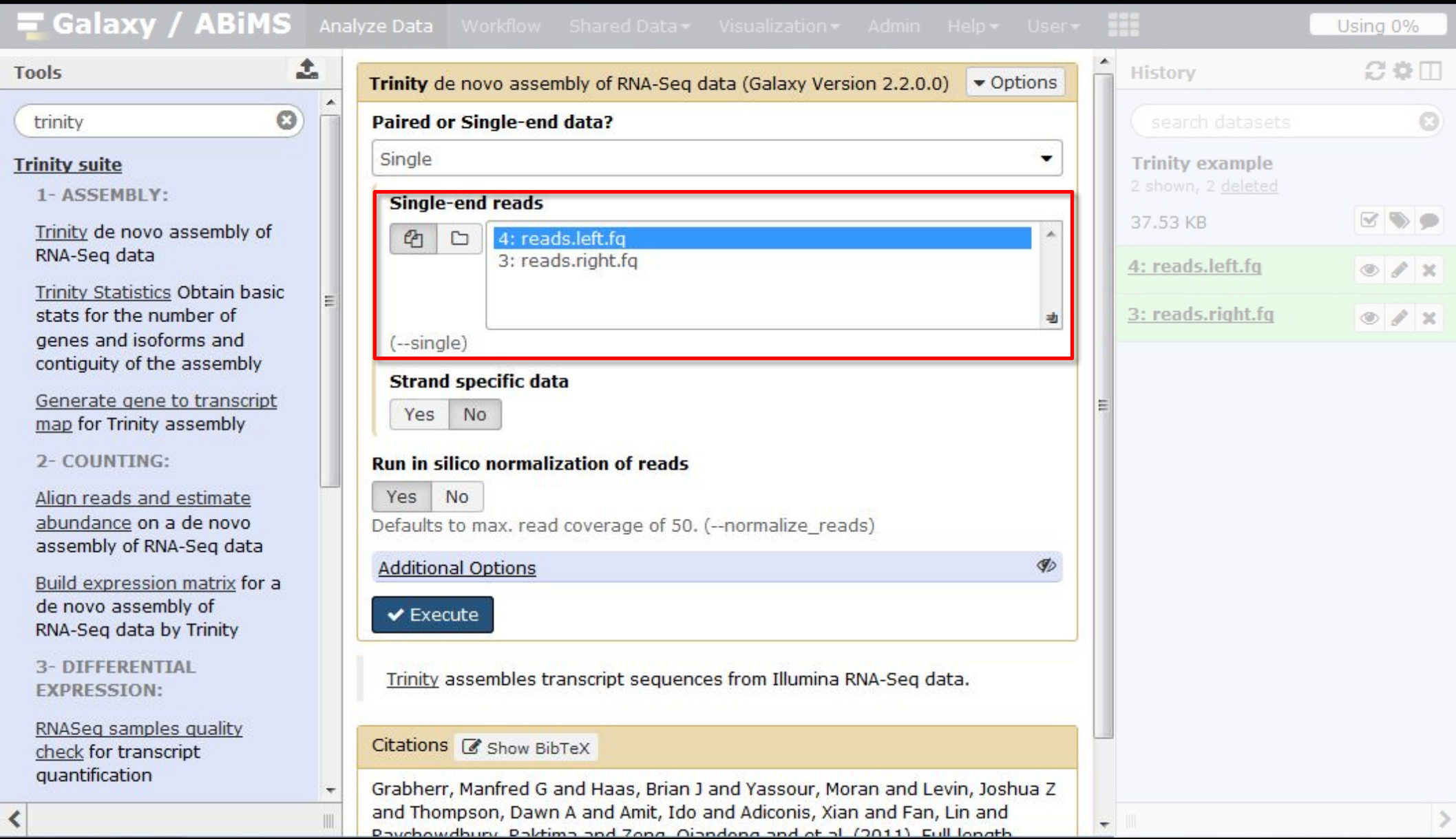
Tools - panel


Galaxy / ABiMS
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 0%
Tools
trinity
Trinity suite
1- ASSEMBLY:
Trinity de novo assembly of RNA-Seq data
Trinity Statistics Obtain basic stats for the number of genes and isoforms and contiguity of the assembly
Generate gene to transcript map for Trinity assembly
2- COUNTING:
Align reads and estimate abundance on a de novo assembly of RNA-Seq data
Build expression matrix for a de novo assembly of RNA-Seq data by Trinity
3- DIFFERENTIAL EXPRESSION:
RNASeq samples quality check for transcript quantification
Welcome to galaxy3.sb-roscoff.fr
Information
For any question or request for tools or account, send an email at support.abims@sb-roscoff.fr
Analyses and Bioinformatics for Marine Science
Station Biologique Roscoff
Changelog
Tutorials
Galaxy is an open, web-based platform for data intensive biomedical research. 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 NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.
History
search datasets
Trinity example
2 shown, 3 deleted
40.02 KB
4: reads.left.fq
3: reads.right.fq

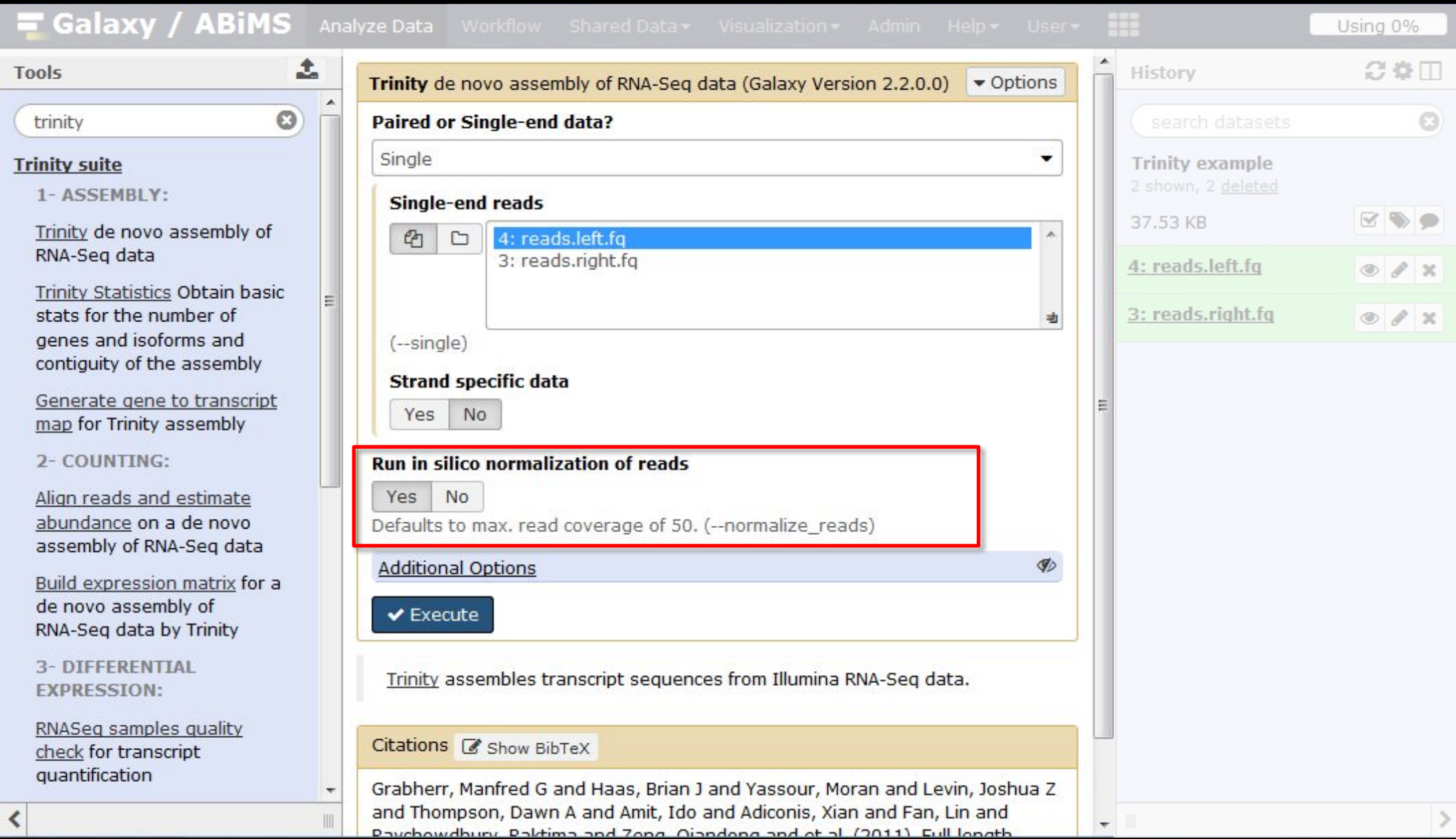
Tools - form


Galaxy / ABiMS
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 0%
Tools
trinity
Trinity suite
1- ASSEMBLY:
Trinity de novo assembly of RNA-Seq data
Trinity Statistics Obtain basic stats for the number of genes and isoforms and contiguity of the assembly
Generate gene to transcript map for Trinity assembly
2- COUNTING:
Align reads and estimate abundance on a de novo assembly of RNA-Seq data
Build expression matrix for a de novo assembly of RNA-Seq data by Trinity
3- DIFFERENTIAL EXPRESSION:
RNASeq samples quality check for transcript quantification
Trinity de novo assembly of RNA-Seq data (Galaxy Version 2.2.0.0)
Options
Paired or Single-end data?
Single
Single-end reads
4: reads.left.fq
3: reads.right.fq
(--single)
Strand specific data
Yes
No
Run in silico normalization of reads
Yes
No
Defaults to max. read coverage of 50. (--normalize_reads)
Additional Options
Execute
Trinity assembles transcript sequences from Illumina RNA-Seq data.
Citations
Show BibTeX
Grabherr, Manfred G and Haas, Brian J and Yassour, Moran and Levin, Joshua Z and Thompson, Dawn A and Amit, Ido and Adiconis, Xian and Fan, Lin and
History
search datasets
Trinity example
2 shown, 2 deleted
37.53 KB
4: reads.left.fq
3: reads.right.fq

Tools - form

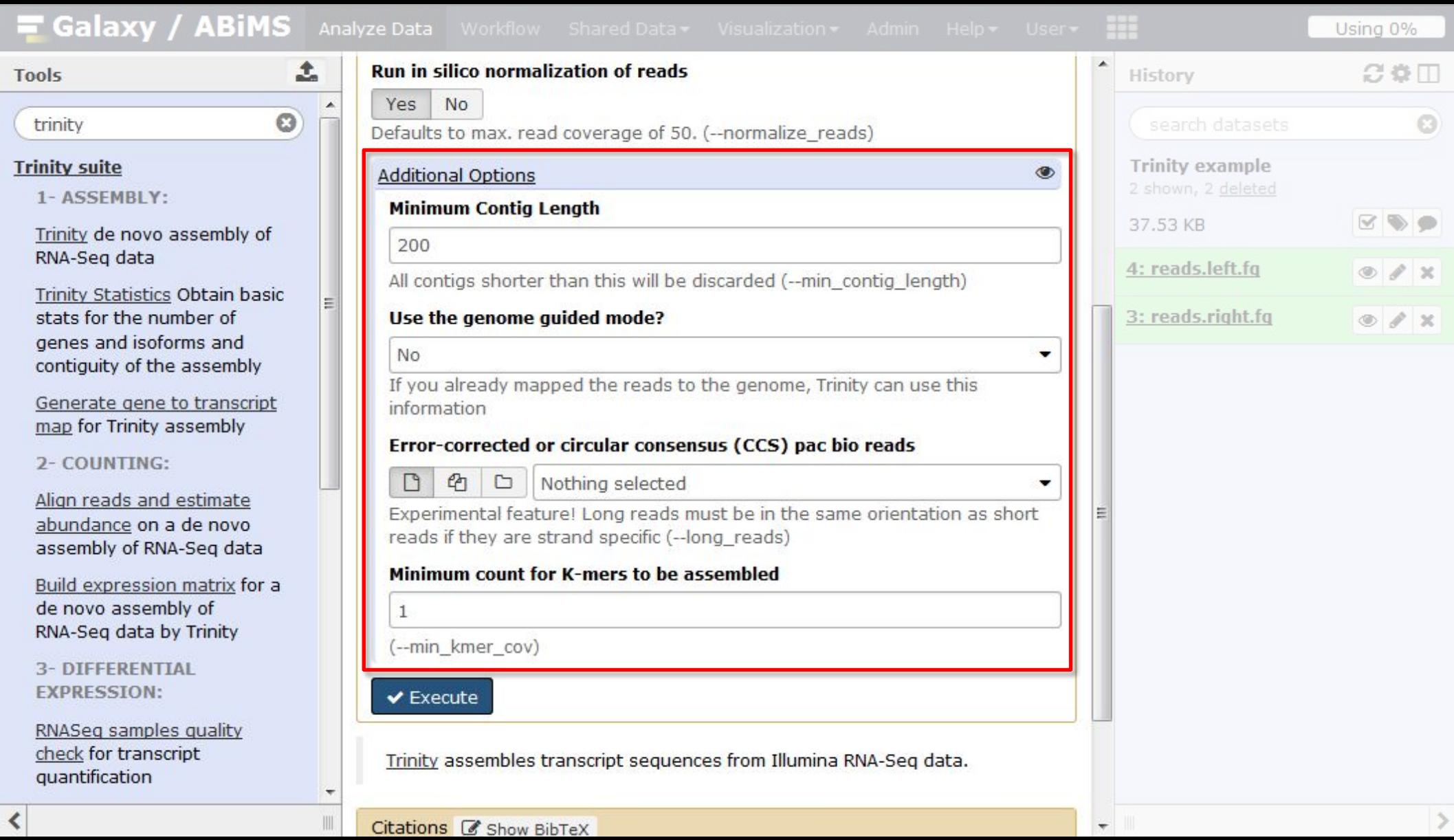

Galaxy / ABiMS
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 0%
Tools
trinity
Trinity suite
1- ASSEMBLY:
Trinity de novo assembly of RNA-Seq data
Trinity Statistics Obtain basic stats for the number of genes and isoforms and contiguity of the assembly
Generate gene to transcript map for Trinity assembly
2- COUNTING:
Align reads and estimate abundance on a de novo assembly of RNA-Seq data
Build expression matrix for a de novo assembly of RNA-Seq data by Trinity
3- DIFFERENTIAL EXPRESSION:
RNASeq samples quality check for transcript quantification
Trinity de novo assembly of RNA-Seq data (Galaxy Version 2.2.0.0)
Options
Paired or Single-end data?
Single
Single-end reads
4: reads.left.fq
3: reads.right.fq
(--single)
Strand specific data
Yes
No
Run in silico normalization of reads
Yes
No
Defaults to max. read coverage of 50. (--normalize_reads)
Additional Options
Execute
Trinity assembles transcript sequences from Illumina RNA-Seq data.
Citations
Show BibTeX
Grabherr, Manfred G and Haas, Brian J and Yassour, Moran and Levin, Joshua Z and Thompson, Dawn A and Amit, Ido and Adiconis, Xian and Fan, Lin and
History
search datasets
Trinity example
2 shown, 2 deleted
37.53 KB
4: reads.left.fq
3: reads.right.fq

Tools - form

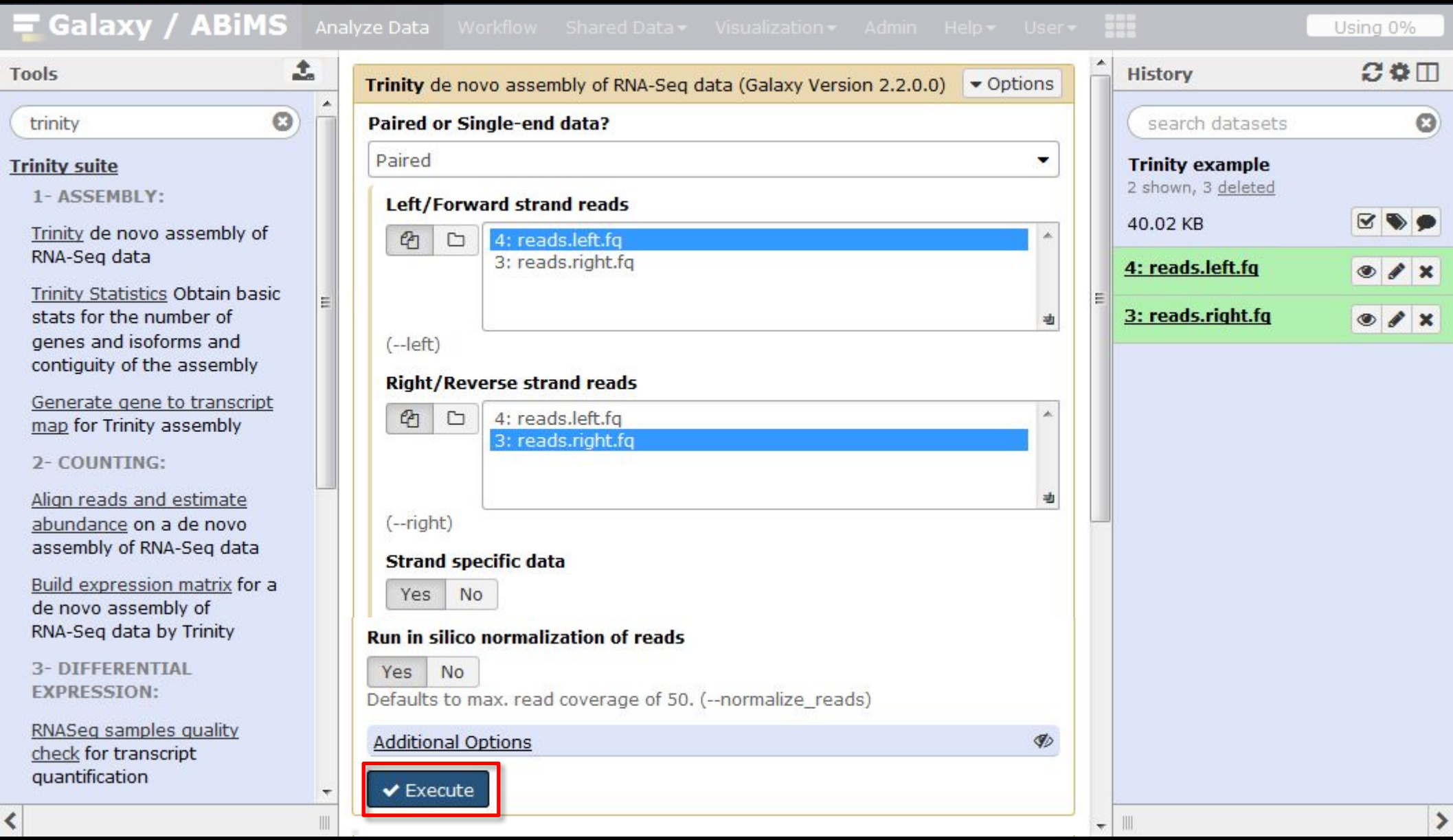

Galaxy / ABiMS
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 0%
Tools
trinity
Trinity suite
1- ASSEMBLY:
Trinity de novo assembly of RNA-Seq data
Trinity Statistics Obtain basic stats for the number of genes and isoforms and contiguity of the assembly
Generate gene to transcript map for Trinity assembly
2- COUNTING:
Align reads and estimate abundance on a de novo assembly of RNA-Seq data
Build expression matrix for a de novo assembly of RNA-Seq data by Trinity
3- DIFFERENTIAL EXPRESSION:
RNASeq samples quality check for transcript quantification
Trinity de novo assembly of RNA-Seq data (Galaxy Version 2.2.0.0)
Options
Paired or Single-end data?
Single
Single-end reads
4: reads.left.fq
3: reads.right.fq
(--single)
Strand specific data
Yes
No
Run in silico normalization of reads
Yes
No
Defaults to max. read coverage of 50. (--normalize_reads)
Additional Options
Execute
Trinity assembles transcript sequences from Illumina RNA-Seq data.
Citations
Show BibTeX
Grabherr, Manfred G and Haas, Brian J and Yassour, Moran and Levin, Joshua Z and Thompson, Dawn A and Amit, Ido and Adiconis, Xian and Fan, Lin and
History
search datasets
Trinity example
2 shown, 2 deleted
37.53 KB
4: reads.left.fq
3: reads.right.fq

Tools - form

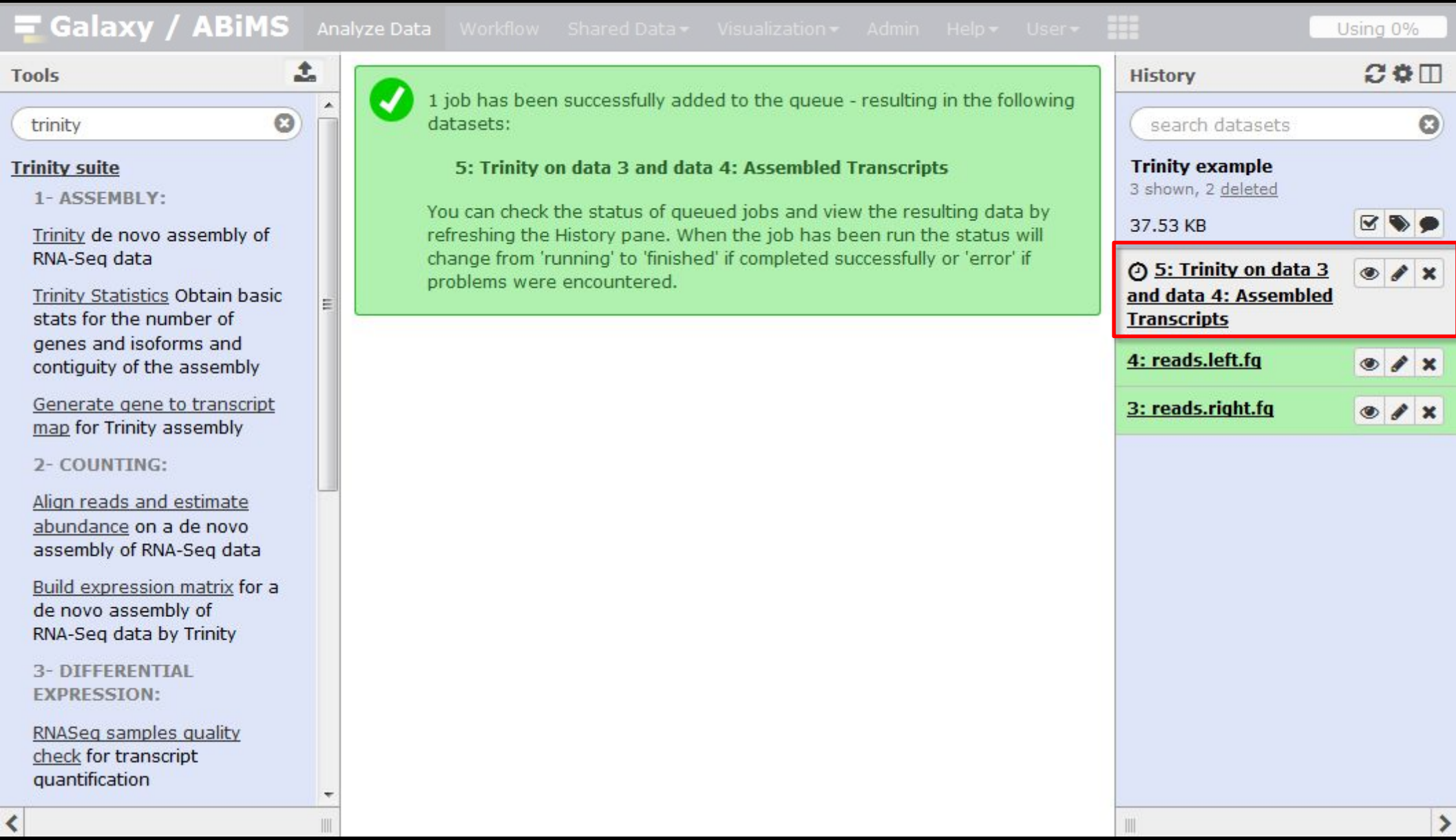
Tools can have some advanced options



Tools - form


Galaxy / ABiMS
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 0%
Tools
trinity
Trinity suite
1- ASSEMBLY:
Trinity de novo assembly of RNA-Seq data
Trinity Statistics Obtain basic stats for the number of genes and isoforms and contiguity of the assembly
Generate gene to transcript map for Trinity assembly
2- COUNTING:
Align reads and estimate abundance on a de novo assembly of RNA-Seq data
Build expression matrix for a de novo assembly of RNA-Seq data by Trinity
3- DIFFERENTIAL EXPRESSION:
RNASeq samples quality check for transcript quantification
Trinity de novo assembly of RNA-Seq data (Galaxy Version 2.2.0.0)
Options
Paired or Single-end data?
Paired
Left/Forward strand reads
4: reads.left.fq
3: reads.right.fq
(--left)
Right/Reverse strand reads
4: reads.left.fq
3: reads.right.fq
(--right)
Strand specific data
Yes
No
Run in silico normalization of reads
Yes
No
Defaults to max. read coverage of 50. (--normalize_reads)
Additional Options
Execute
History
search datasets
Trinity example
2 shown, 3 deleted
40.02 KB
4: reads.left.fq
3: reads.right.fq

Tools - form


Galaxy / ABiMS
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 0%
Tools
trinity
Trinity suite
1- ASSEMBLY:
Trinity de novo assembly of RNA-Seq data
Trinity Statistics Obtain basic stats for the number of genes and isoforms and contiguity of the assembly
Generate gene to transcript map for Trinity assembly
2- COUNTING:
Align reads and estimate abundance on a de novo assembly of RNA-Seq data
Build expression matrix for a de novo assembly of RNA-Seq data by Trinity
3- DIFFERENTIAL EXPRESSION:
RNASeq samples quality check for transcript quantification
1 job has been successfully added to the queue - resulting in the following datasets:
5: Trinity on data 3 and data 4: Assembled Transcripts
You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.
History
search datasets
Trinity example
3 shown, 2 deleted
37.53 KB
5: Trinity on data 3 and data 4: Assembled Transcripts
4: reads.left.fq
3: reads.right.fq

Tools - Job status

Job is waiting to run

= the job is in the scheduler « queue »

Duration time of this status depends on the amount of actual queued jobs and on the requested number of processors

Tools - Job status

Job is currently running

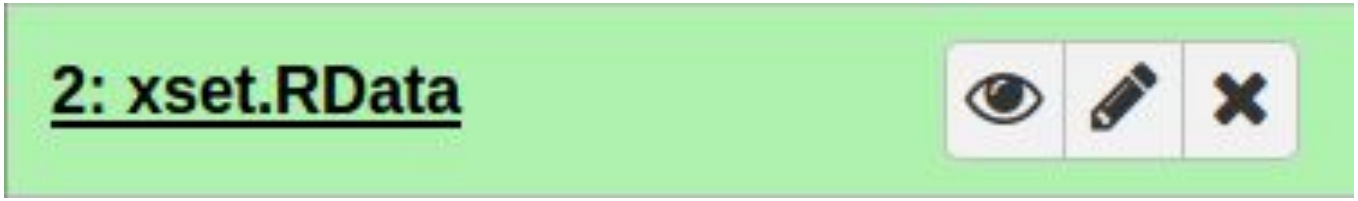
= the job is being executed on the computing cluster

Duration time of this status depends on the job's attributes and the computing resources allocated.

Some programs are executed with several processors (using 4, 8 or 16 Gb of RAM).

And others are mono-threaded ☹

Tools - Job status



Job is finished and status is OK

But warnings or errors can be hidden behind!

Tools - Job status

Job is finished but with an error status

= the program sends an error

The error is often explained by the program but sometimes … not.

Tools - Job status

Job is finished but with an error status

= the program sends an error

Possible causes of error :

- ~~The user~~ :P
- Bad usage : input file, format or option
- Bad integration of the program into Galaxy … sorry :/
- Non anticipated crash of the program

Tools - Handle errors

Tools - Handle errors

Galaxy / ABiMS | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 0%

Tools

search tools

Get Data
Send Data
Collection Operations

COMMON TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Statistics
Graph/Display Data
Fasta Fastq Manipulation

COMMON NGS TOOLS
NGS:Samtools
NGS:Mapping
NGS:Bedtools
NGS:Picard Tools

SEARCHING TOOLS
Diamond

Dataset generation errors

Dataset 48: group2_count2.txt

The Galaxy framework encountered the following error while attempting to run the tool:

```
Traceback (most recent call last):
  File "/w/galaxy/galaxy3/galaxy/lib/galaxy/jobs/runners/local.py", line
    stdout_file.close()
  File "/opt/python/lib/python2.7/tempfile.py", line 403, in close
    self.unlink(self.name)
OSError: [Errno 2] No such file or directory: '/w/gala
                                                    tabase
```

Tool execution generated the following error message:

```
failure running job
```

Report this error to the local Galaxy administrators

Usually the local Galaxy administrators regularly review errors that occur on the server. However, if you would like to provide additional information (such as what you were trying to do when the error occurred) and a contact e-mail address, we will be better able to investigate your problem and get back to you.

Error Report

Your email

e.gueguen@sb-roscoff.fr

galaxy3.sb-roscoff.fr/dataset/errors?id=53eb0ef5c8056a28

History

search datasets

eba 2016 sartools
28 shown, 14 deleted

1.59 MB

48: group2_count2.txt

tool error
An error occurred with this dataset:

failure running job

View or report this error

gene1	1353
gene10	72
gene100	496
gene1000	50

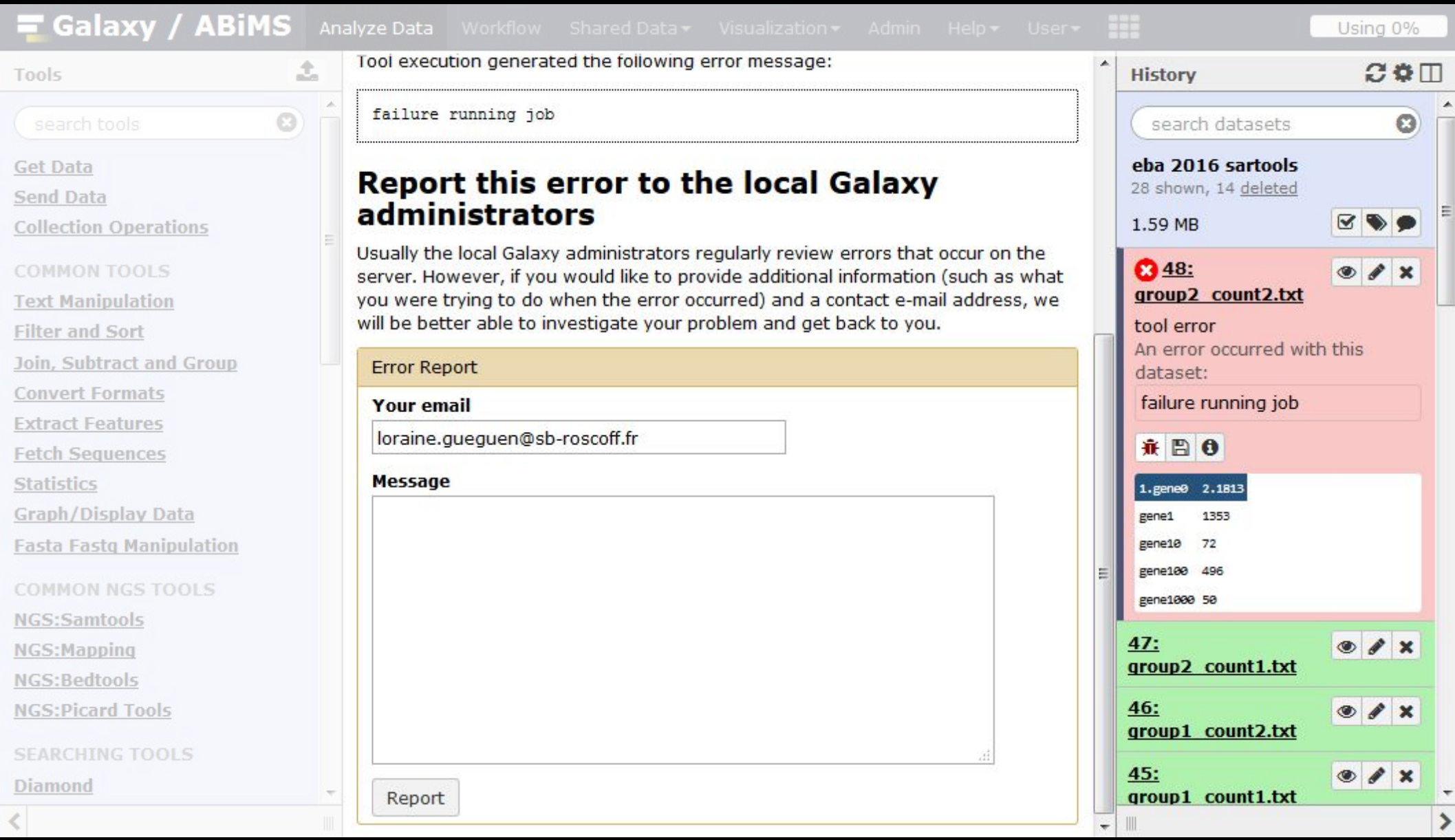
47: group2_count1.txt

46: group1_count2.txt

45: group1_count1.txt

Tools - Handle errors

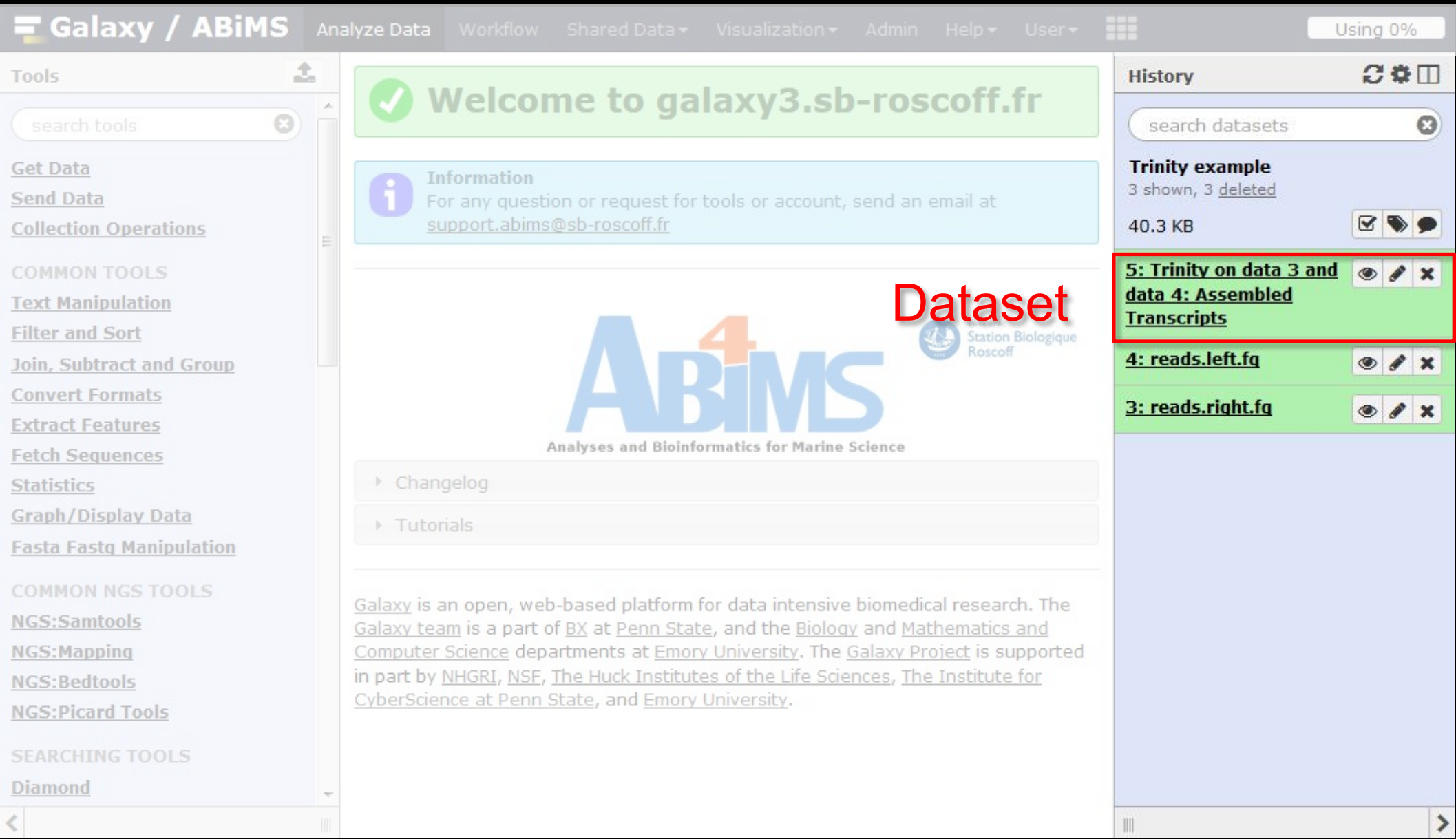
Sent to the support team



DATASET

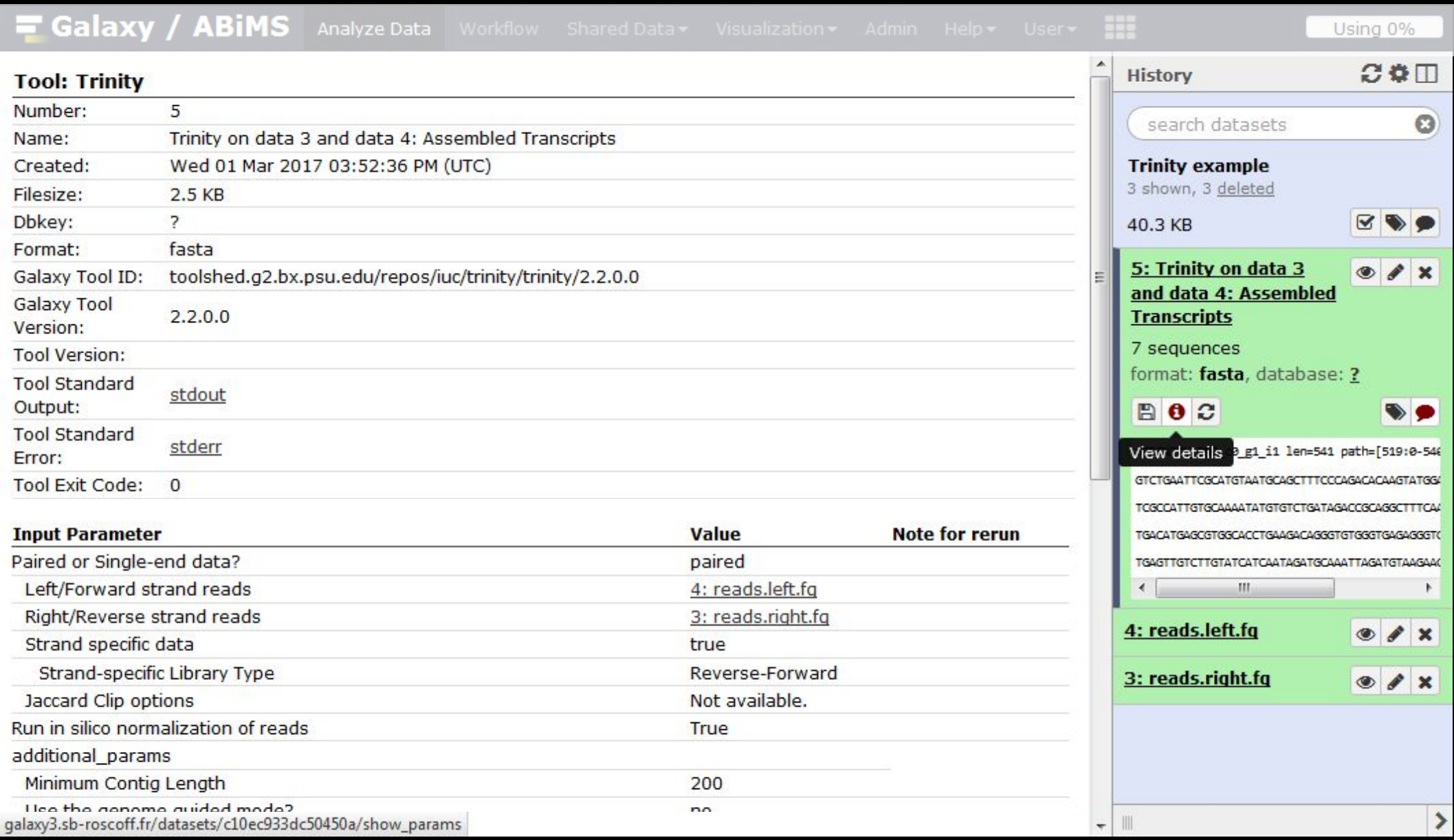
Dataset

Both inputs and outputs



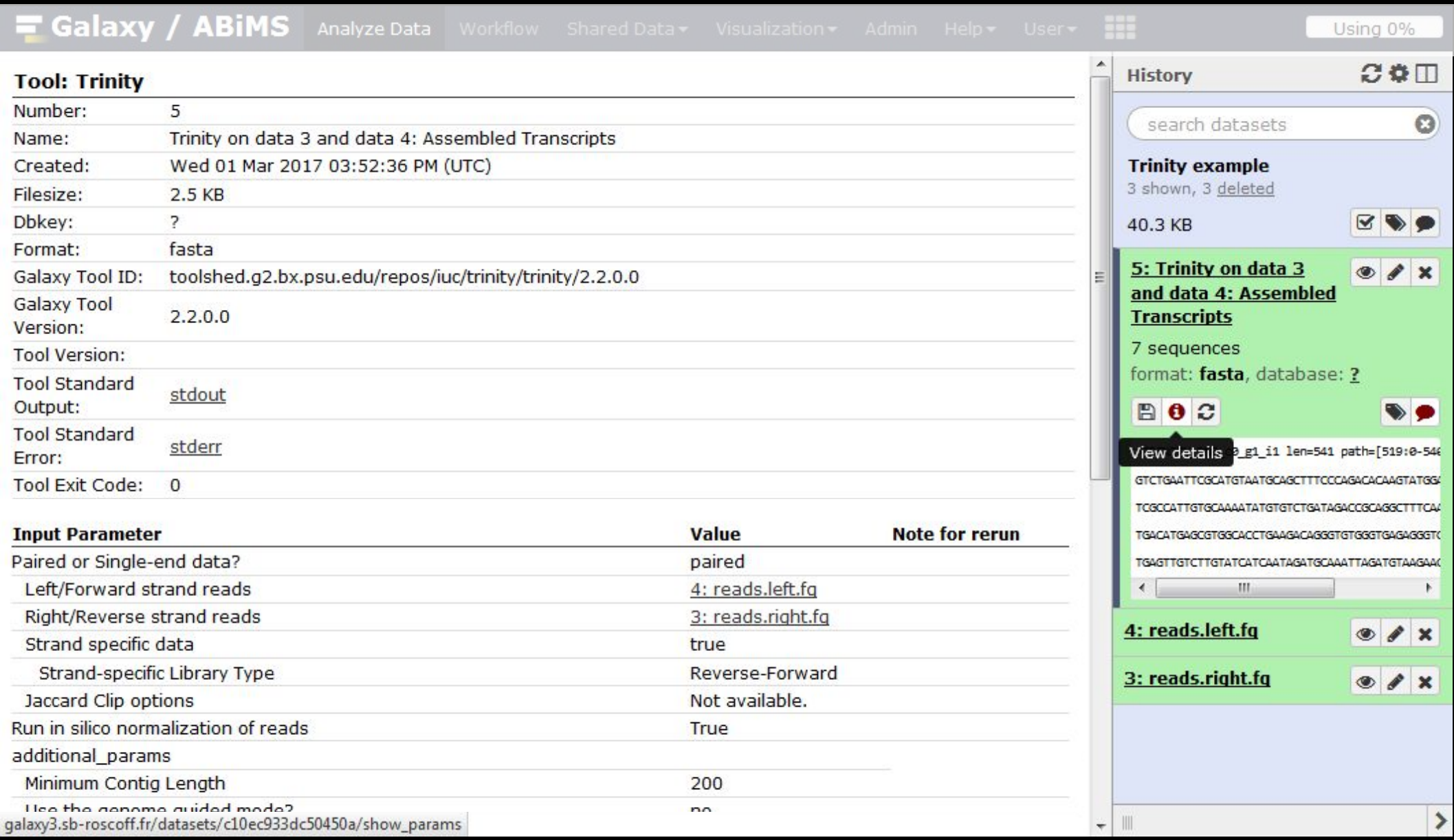
Dataset

Informations



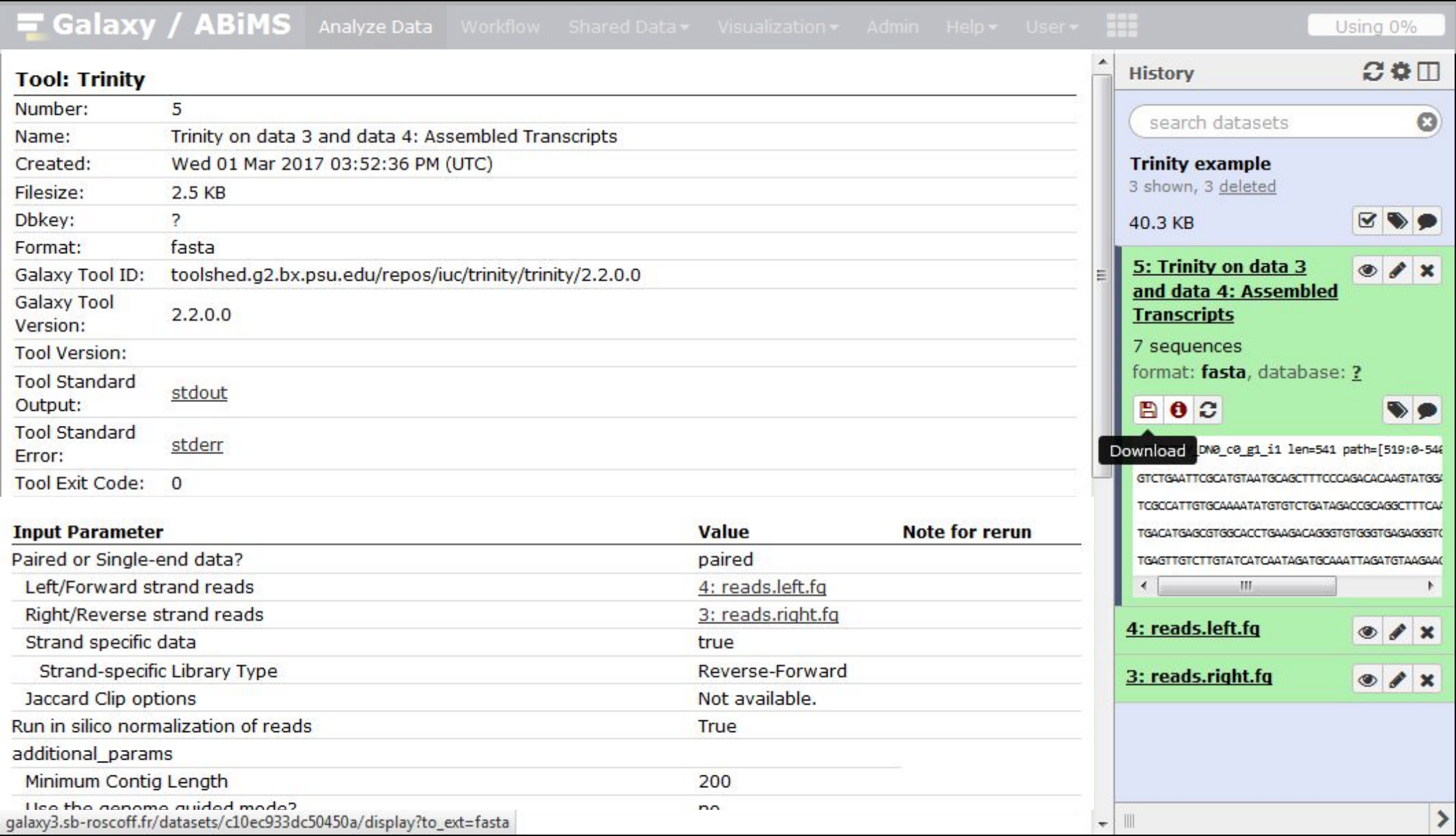
Dataset

Informations



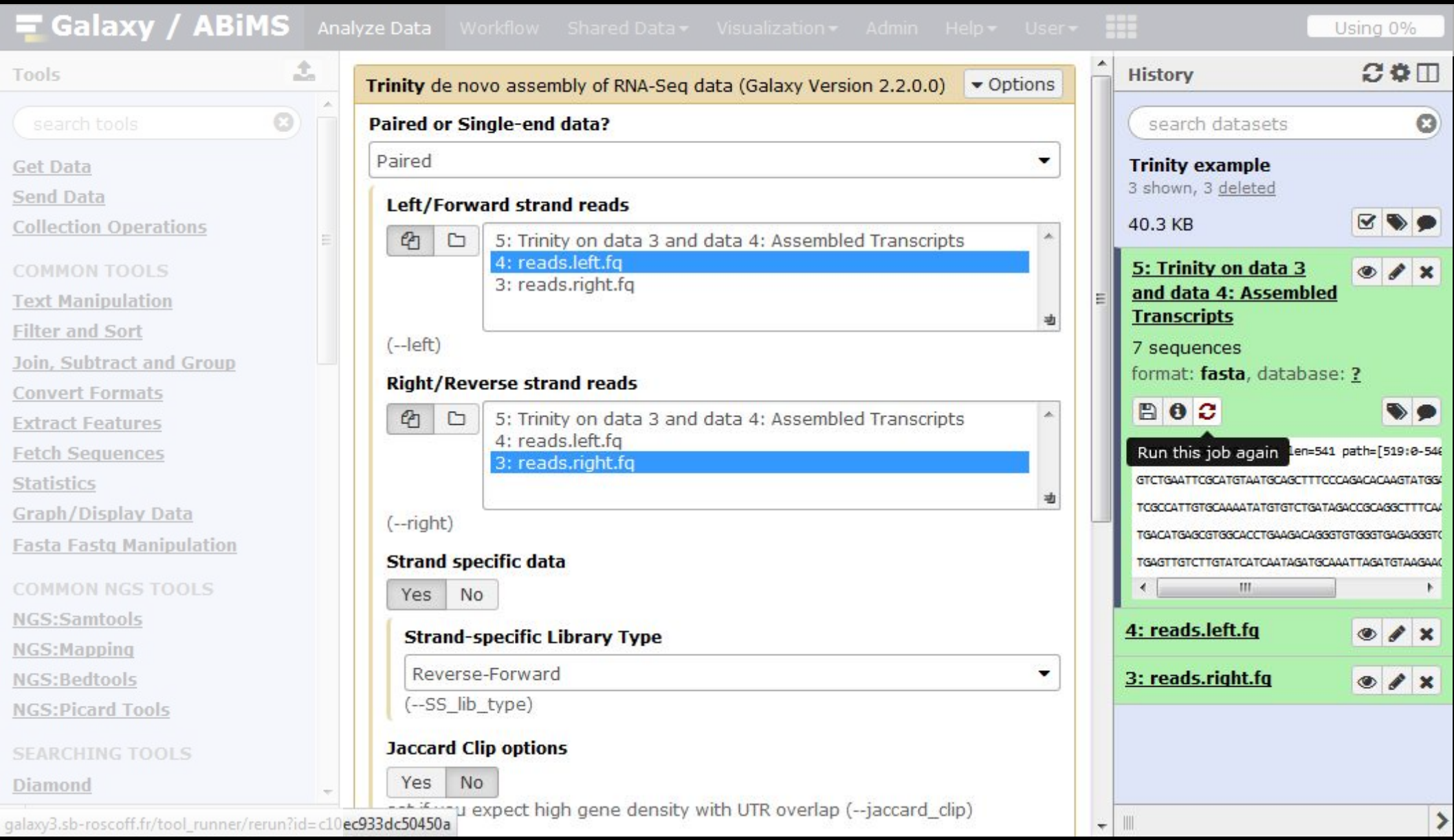
Dataset

Download



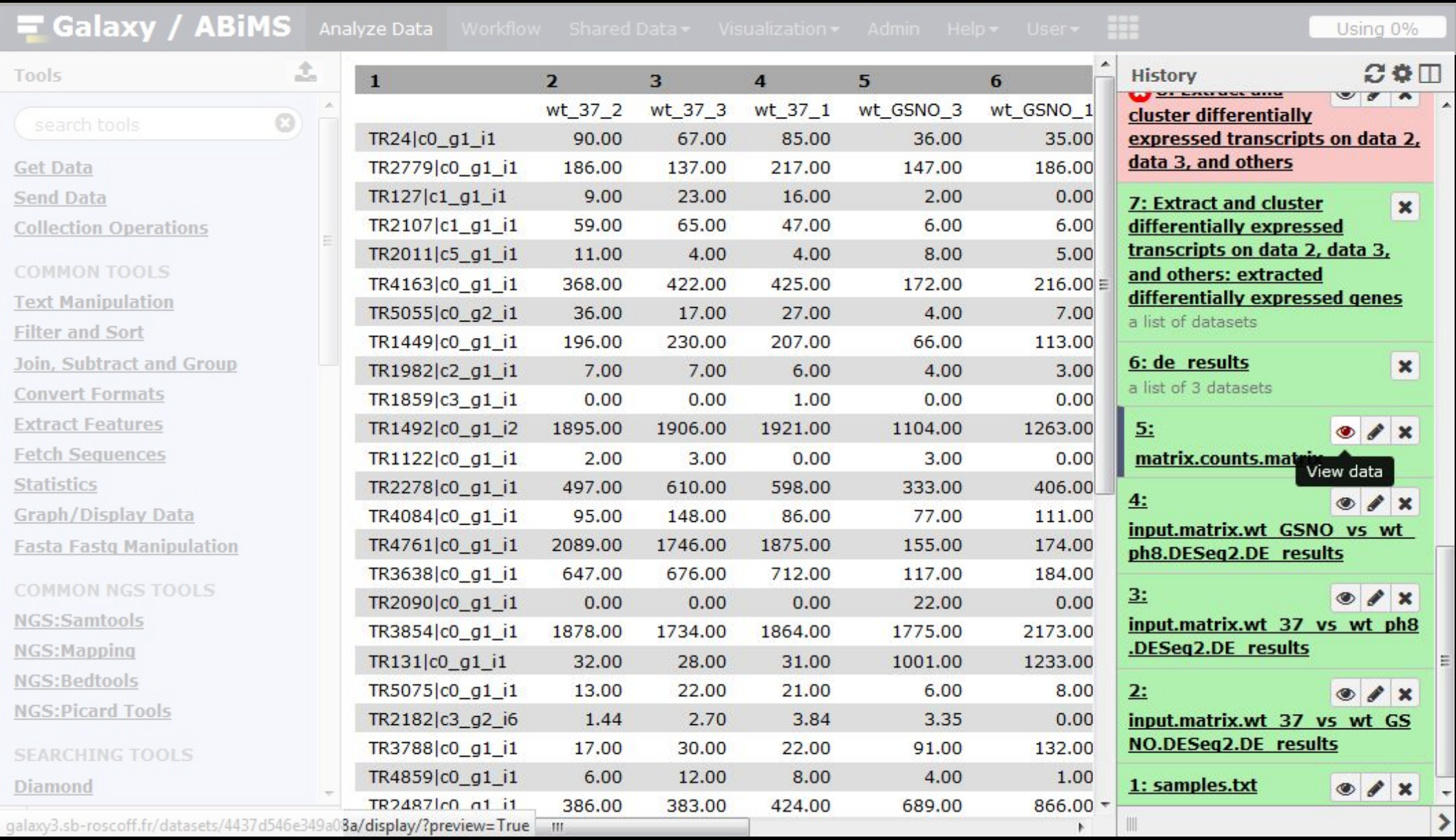
Dataset

Re-run a job



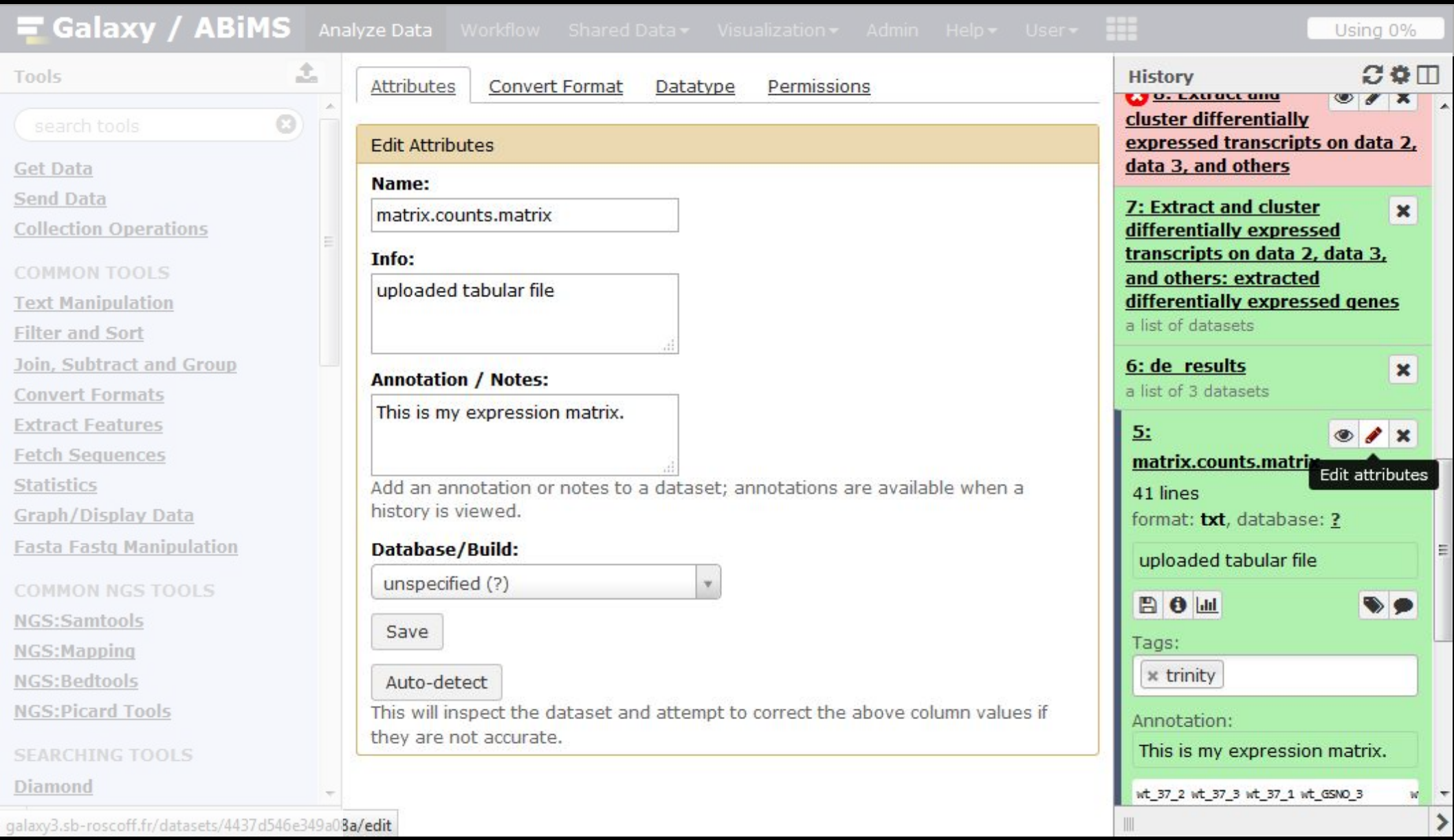
Dataset

Dataset display : text, tabular, pdf, picture, html …



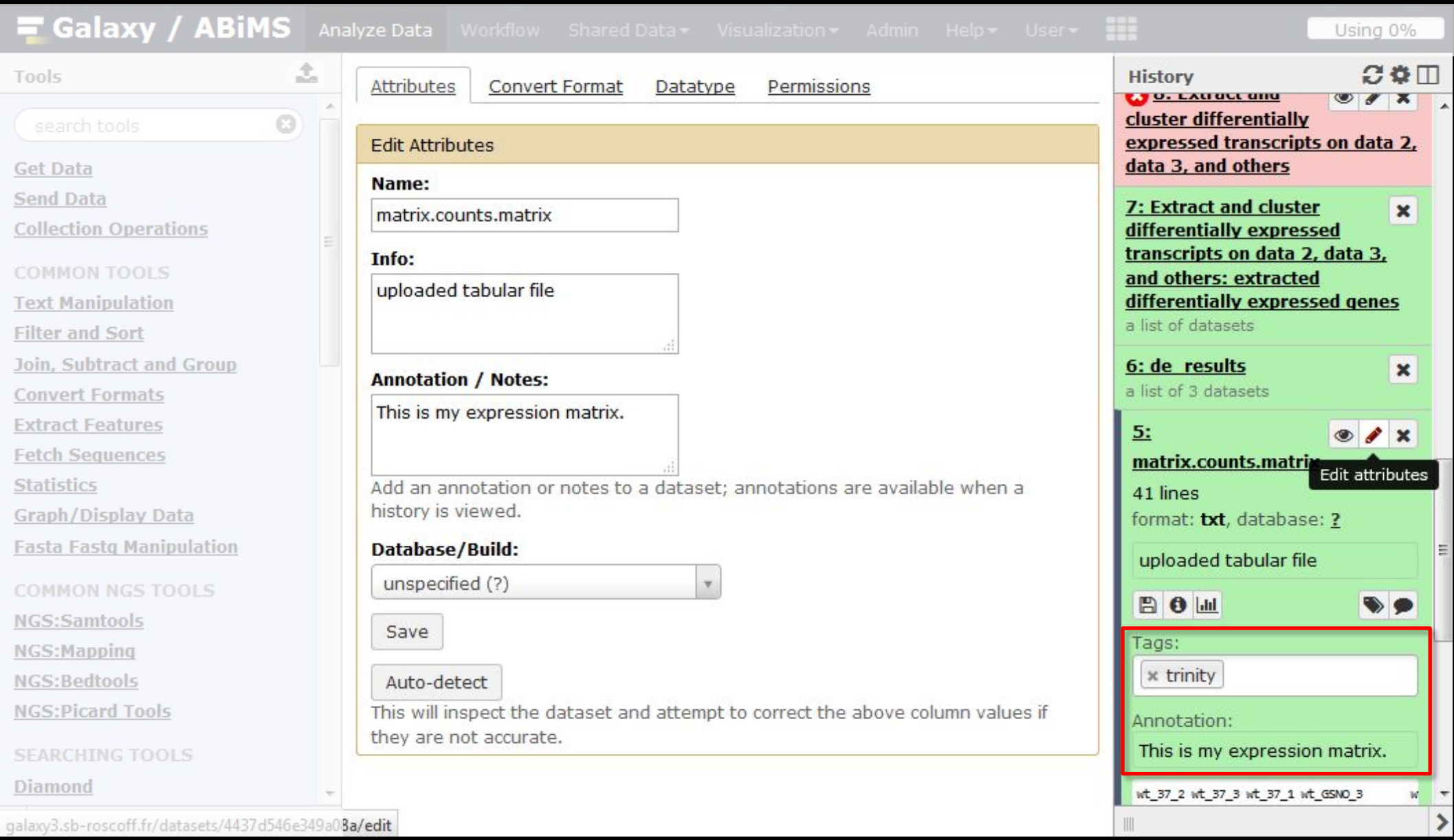
Dataset

Renaming and annotation



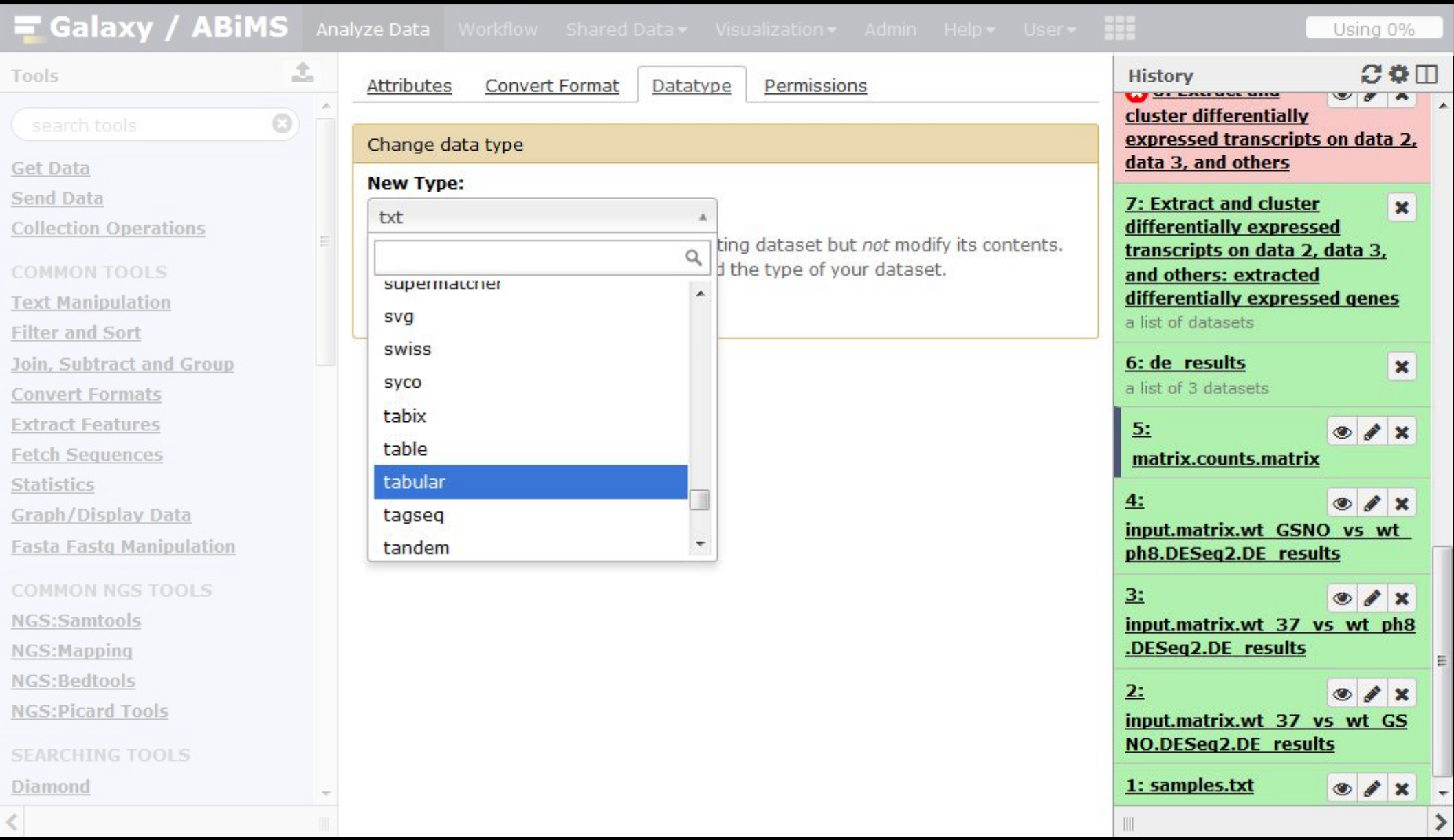
Dataset

Renaming and annotation

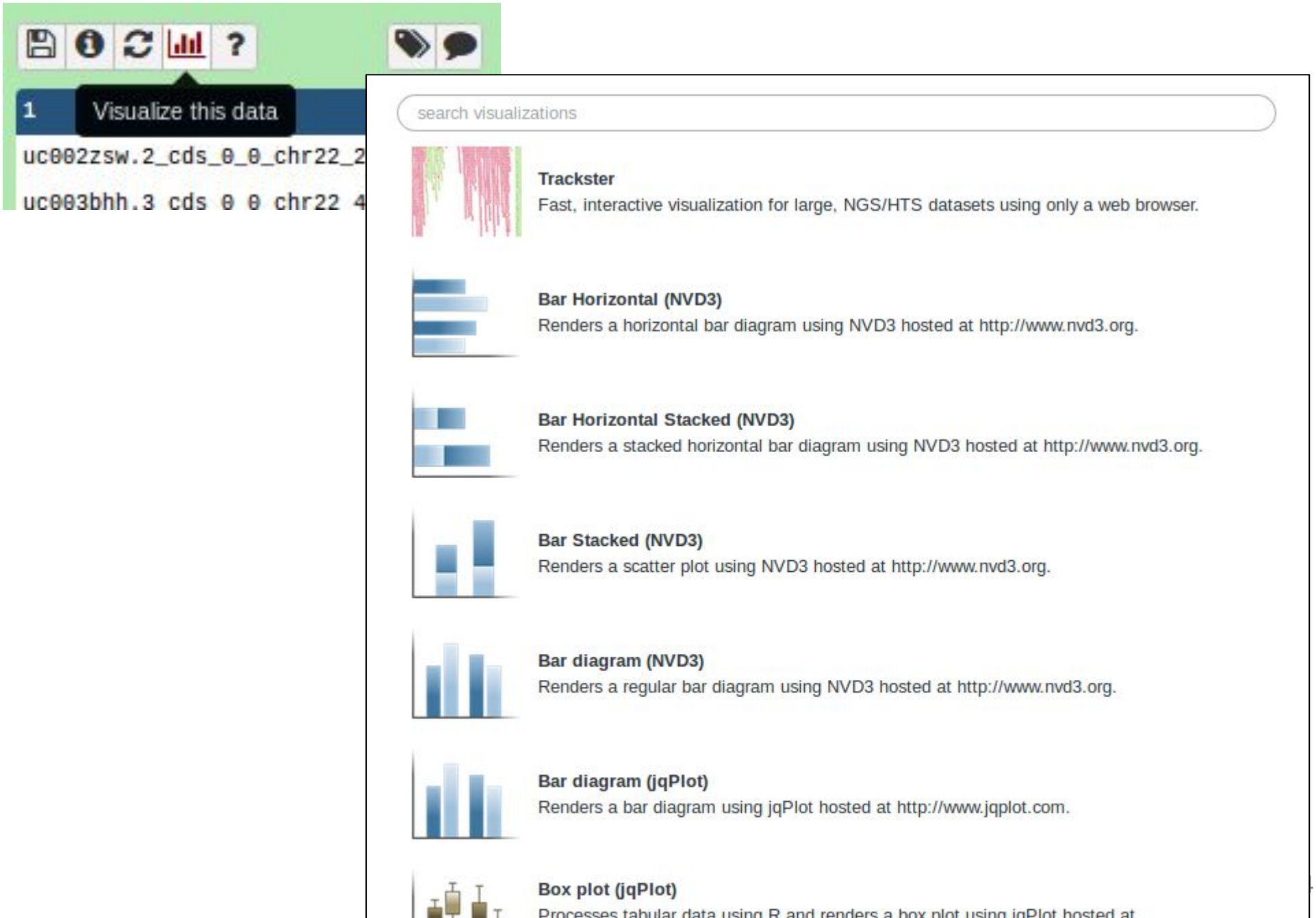


Dataset

Change the Datatype of the Dataset



Dataset


Visualize this data
1
uc002zsw.2_cds_0_0_chr22_2
uc003bhh.3 cds 0 0 chr22 4
search visualizations
Trackster
Fast, interactive visualization for large, NGS/HTS datasets using only a web browser.
Bar Horizontal (NVD3)
Renders a horizontal bar diagram using NVD3 hosted at http://www.nvd3.org.
Bar Horizontal Stacked (NVD3)
Renders a stacked horizontal bar diagram using NVD3 hosted at http://www.nvd3.org.
Bar Stacked (NVD3)
Renders a scatter plot using NVD3 hosted at http://www.nvd3.org.
Bar diagram (NVD3)
Renders a regular bar diagram using NVD3 hosted at http://www.nvd3.org.
Bar diagram (jqPlot)
Renders a bar diagram using jqPlot hosted at http://www.jqplot.com.
Box plot (jqPlot)

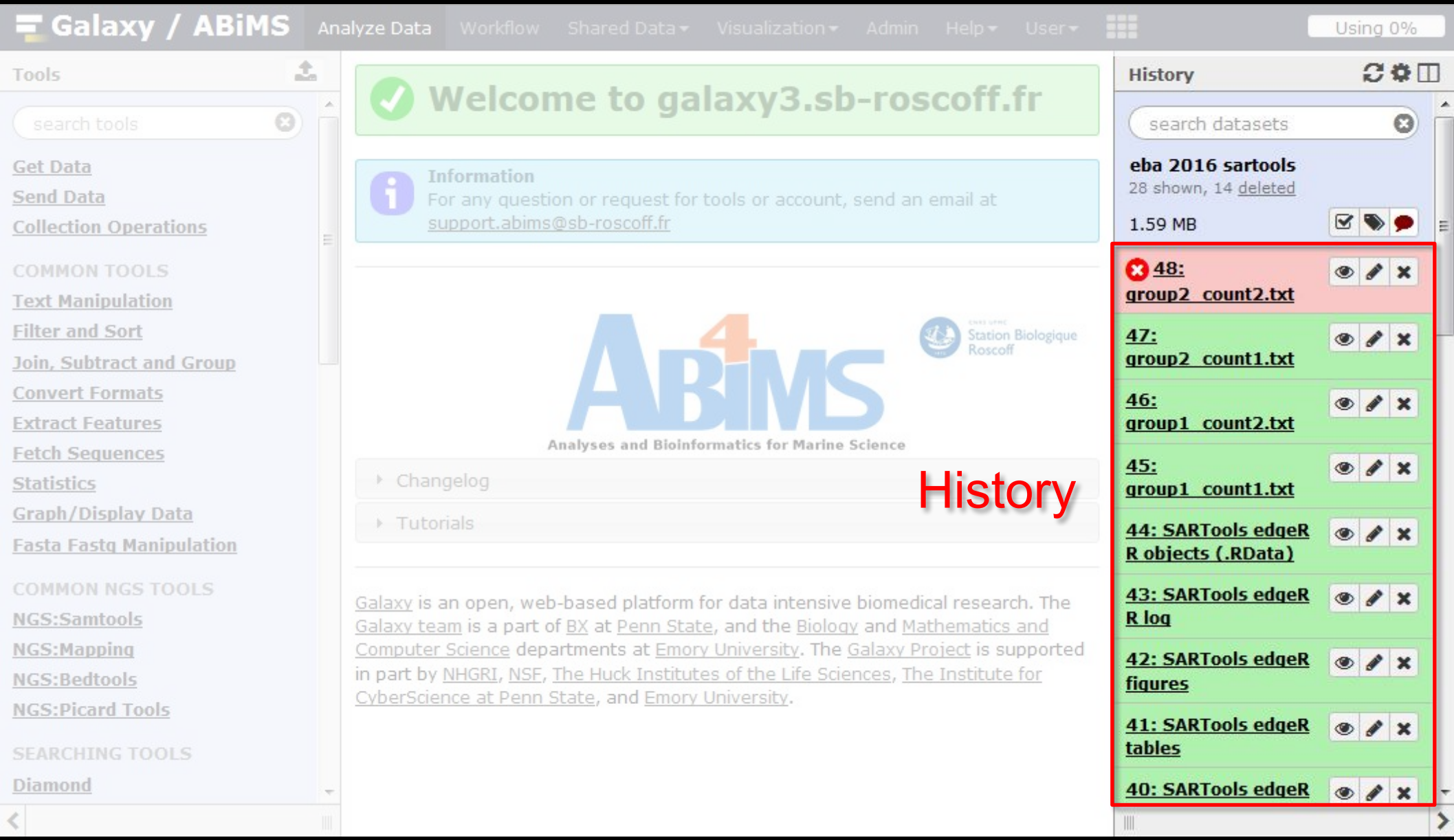
Dataset


7: Select first on data 6
5 lines
format: tabular, database: hg38
Tool Help
1 2
uc010gqp.3_cds_0_0_chr22_15690078_f 63
uc011agd.3_cds_0_0_chr22_15528159_f 48
uc062bek.1_cds_0_0_chr22_15690246_f 46
uc003bhh.4_cds_0_0_chr22_46256561_r 30
uc062bej.1_cds_1_0_chr22_15690426_f 26
Tool help for Select first
What it does
This tool outputs specified number of lines from the beginning of a dataset
Example
Selecting 2 lines from this:
chr7 56632 56652 D17003_CTCF_R6
chr7 56736 56756 D17003_CTCF_R7
chr7 56761 56781 D17003_CTCF_R4
chr7 56772 56792 D17003_CTCF_R7
chr7 56775 56795 D17003_CTCF_R4
will produce:
chr7 56632 56652 D17003_CTCF_R6
chr7 56736 56756 D17003_CTCF_R7

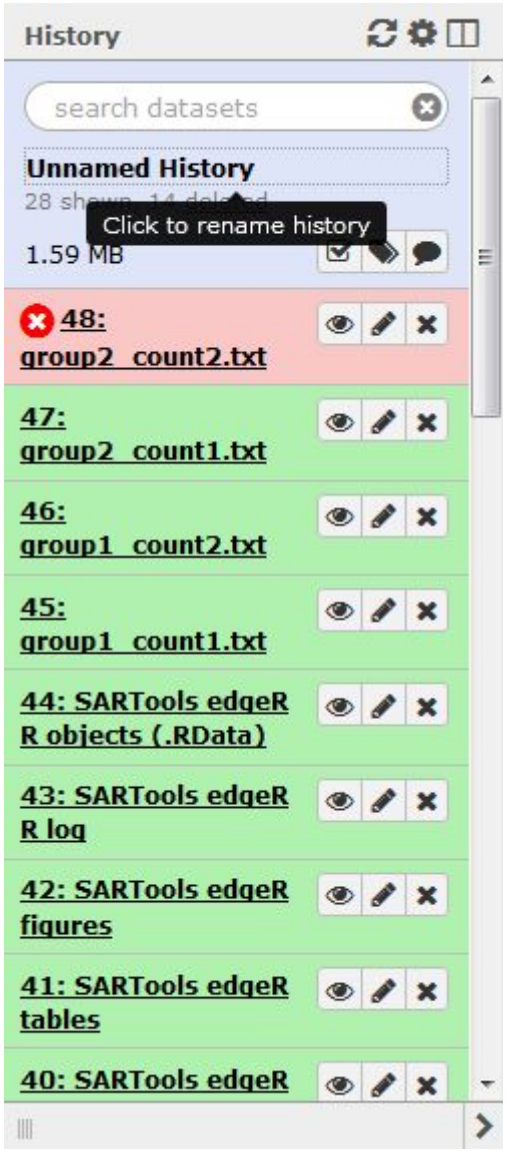
HISTORY

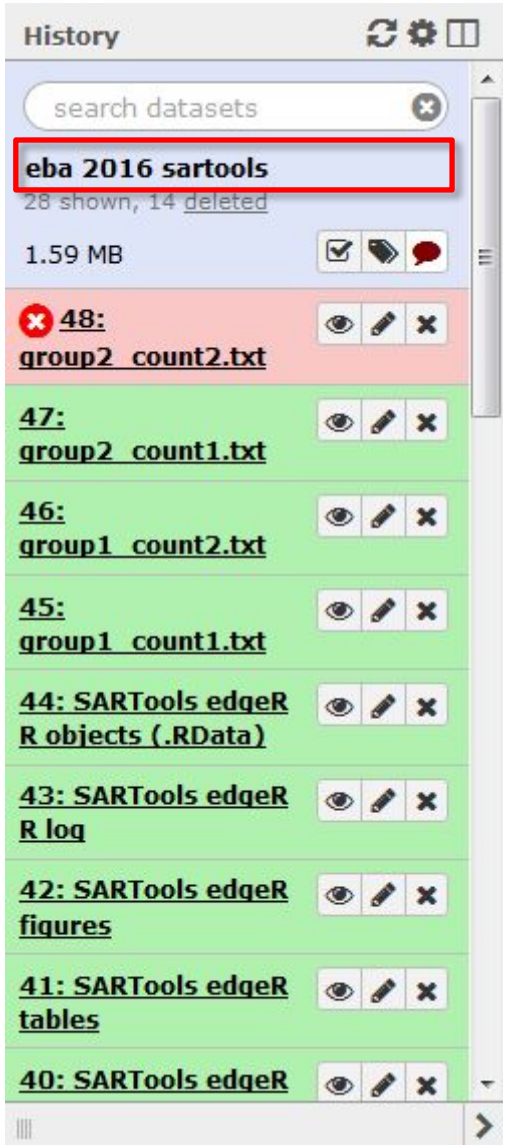
History panel

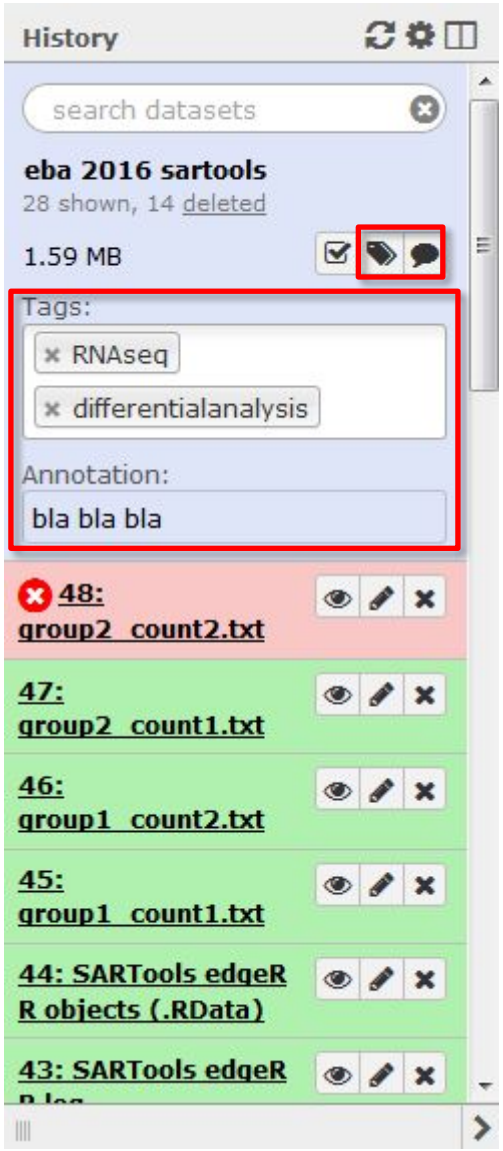
Both inputs and outputs



History panel renaming and annotation

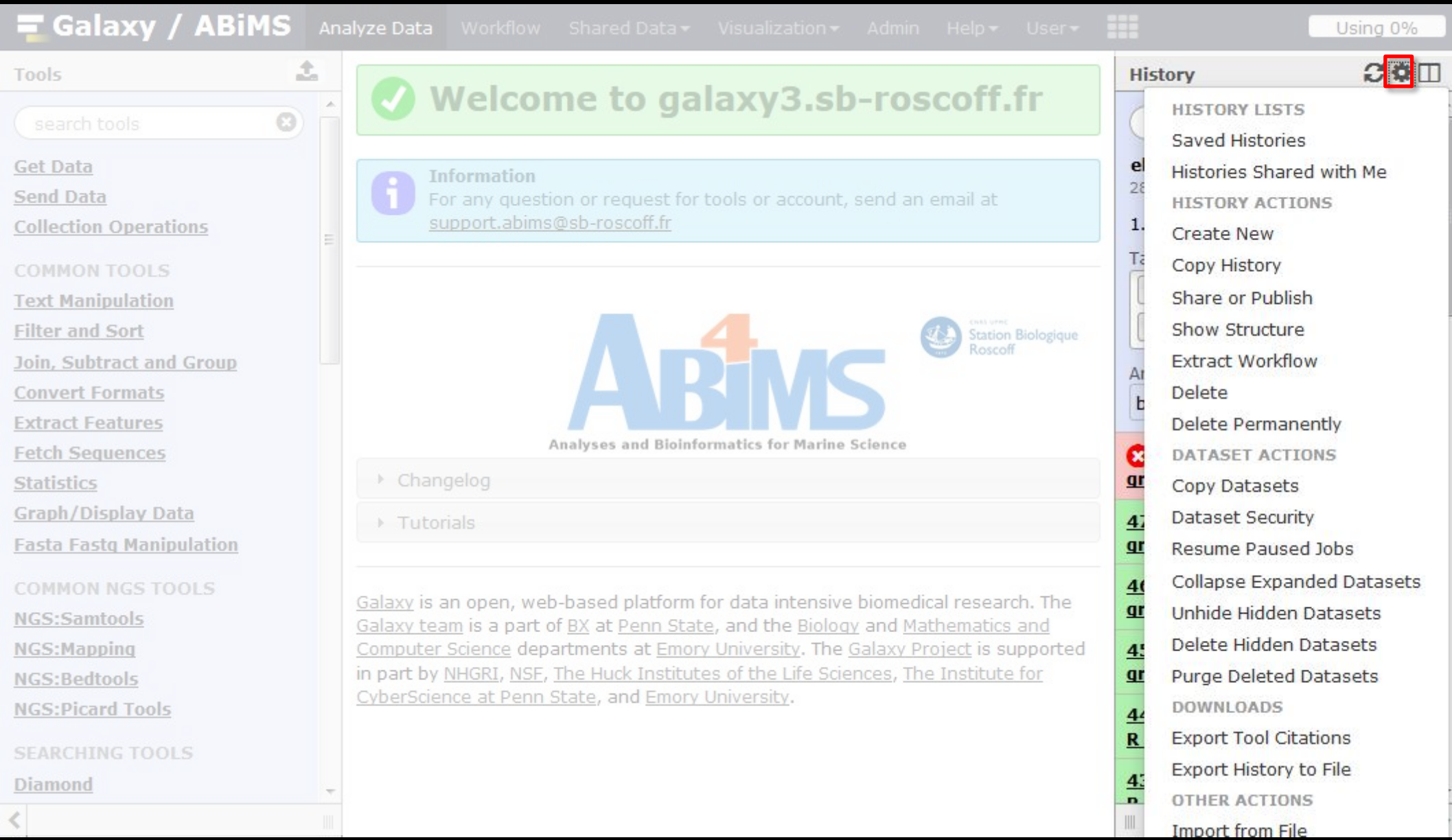






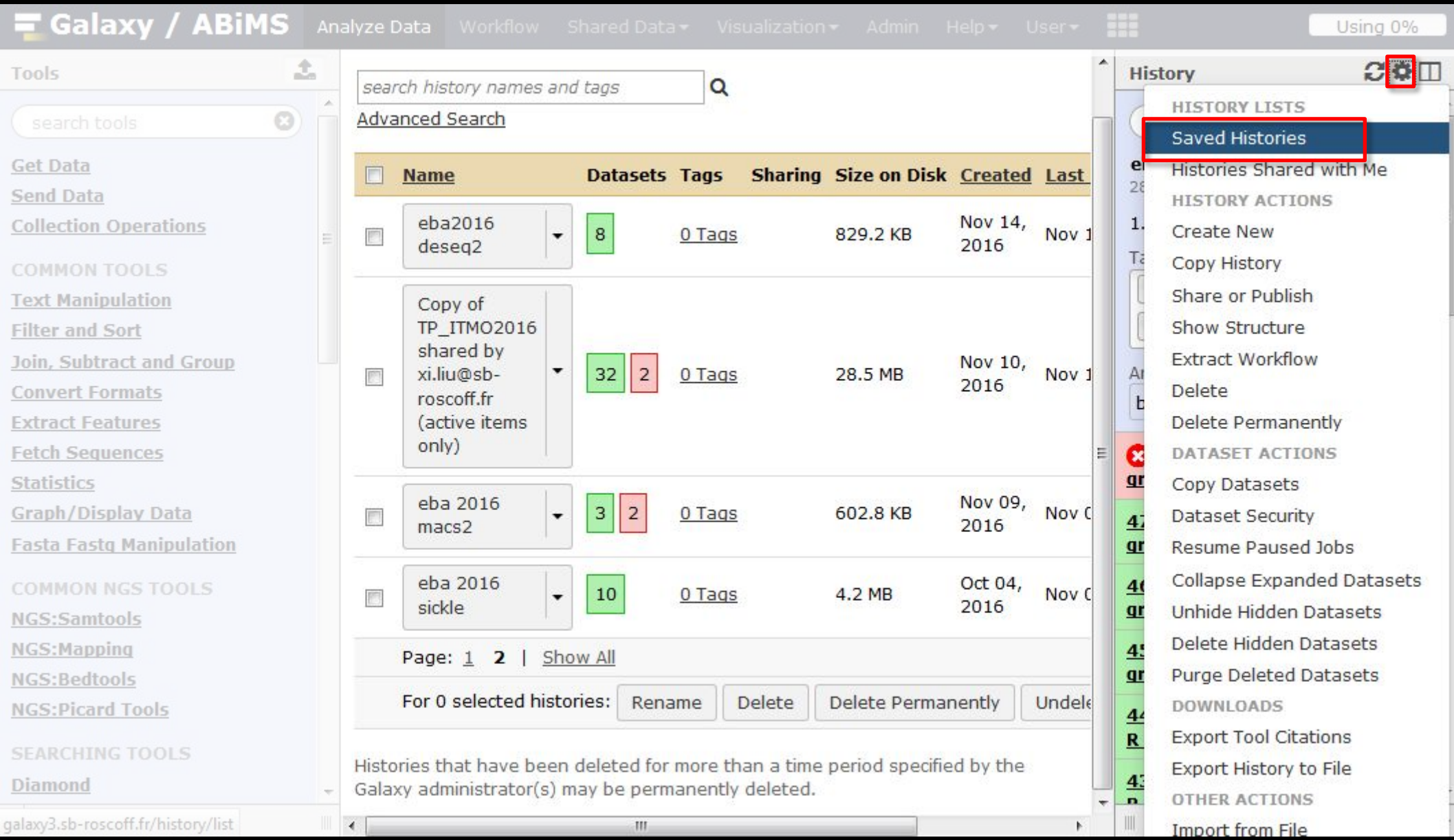
History panel

History menu: Create new, Rename, Delete, Delete Permanently



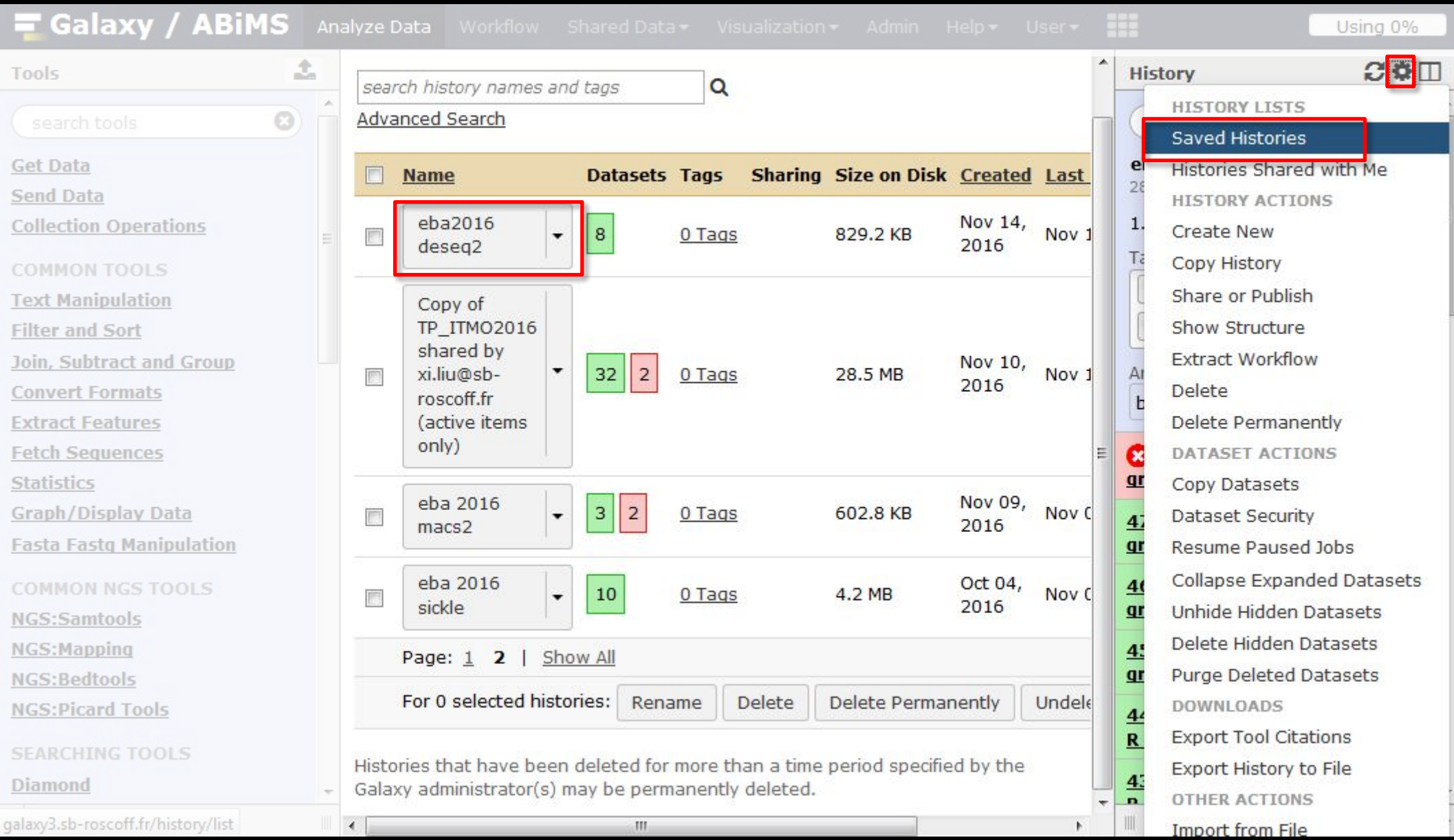
History panel

Saved histories



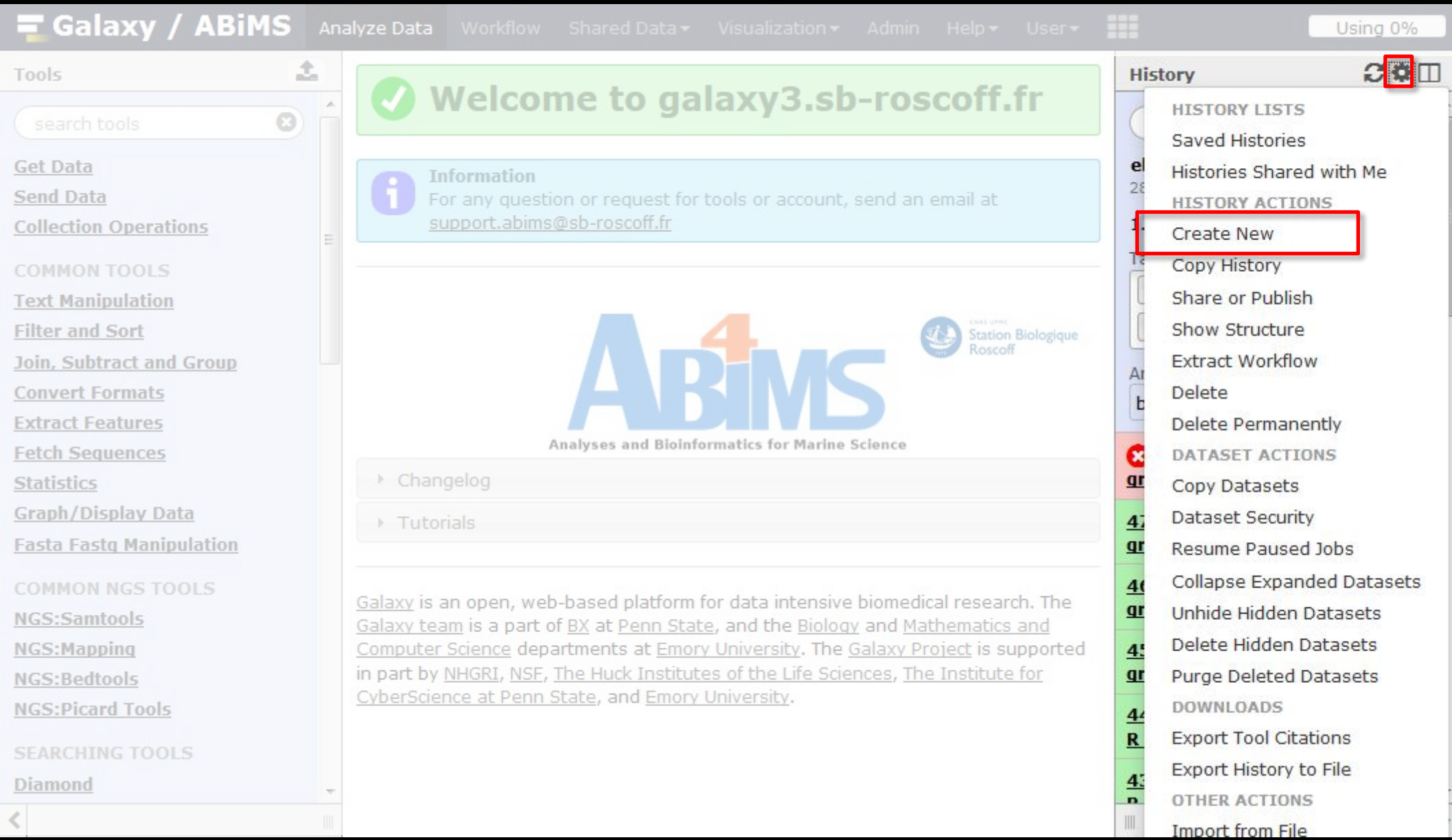
History panel

Saved histories: Switch histories



History panel

History menu: Create new, Rename, Delete, Delete Permanently



History panel

History panel

History panel

History panel

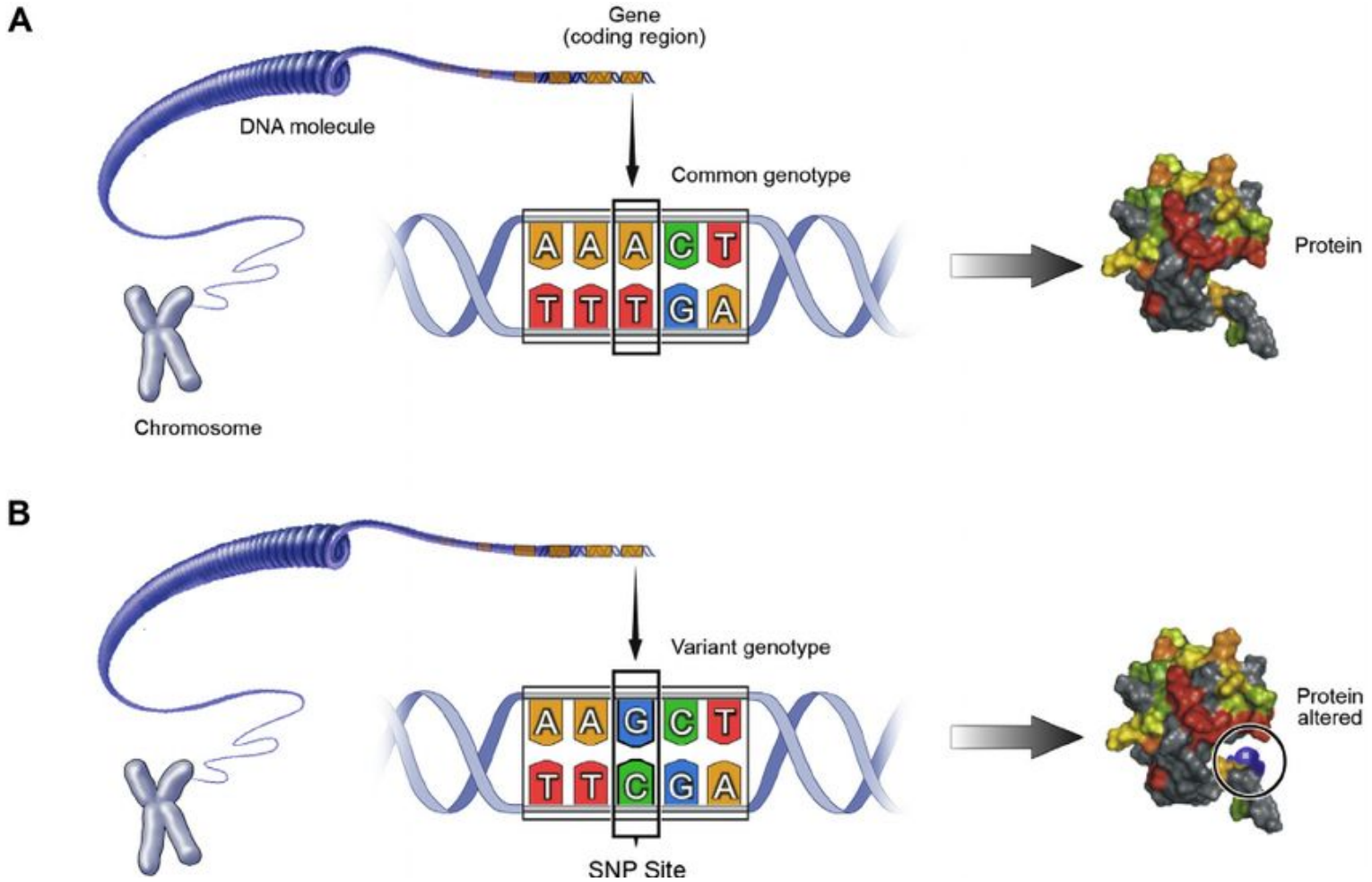
History panel

History panel

Hands-on

TOOLS

Which coding exon has the highest number of single nucleotide polymorphisms on chromosome 22?



DOI: 10.1016/j.jand.2013.12.001

Tools – Hands-on

Which coding exon has the highest number of single nucleotide polymorphisms on chromosome 22?

exons.bed

Chrom	Start	End	Name	Score	Strand
chr22	16258185	16258303	uc002zlh.1_cds_1_0_chr22_16258186_r	0	-
chr22	16266928	16267095	uc002zlh.1_cds_2_0_chr22_16266929_r	0	-
chr22	16268136	16268181	uc002zlh.1_cds_3_0_chr22_16268137_r	0	-

snps.bed

Chrom	Start	End	Name	Score	Strand
chr22	16266919	16266920	rs757764551	0	+
chr22	16266920	16266920	rs375488594	0	+
chr22	16267014	16267015	rs544633418	0	+
chr22	16267029	16267030	rs563306354	0	+
chr22	16267081	16267082	rs568292779	0	+

Tools – Hands-on

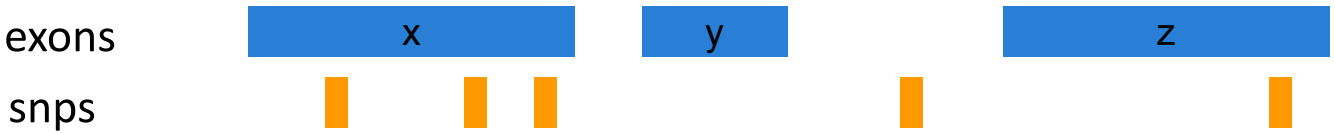
Which coding exon has the highest number of single nucleotide polymorphisms on chromosome 22?

exons.bed

Chrom	Start	End	Name	Score	Strand
chr22	16258185	16258303	uc002zlh.1_cds_1_0_chr22_16258186_r	0	-
chr22	16266928	16267095	uc002zlh.1_cds_2_0_chr22_16266929_r	0	-
chr22	16268136	16268181	uc002zlh.1_cds_3_0_chr22_16268137_r	0	-

snps.bed

Chrom	Start	End	Name	Score	Strand
chr22	16266919	16266920	rs757764551	0	+
chr22	16266920	16266920	rs375488594	0	+
chr22	16267014	16267015	rs544633418	0	+
chr22	16267029	16267030	rs563306354	0	+
chr22	16267081	16267082	rs568292779	0	+



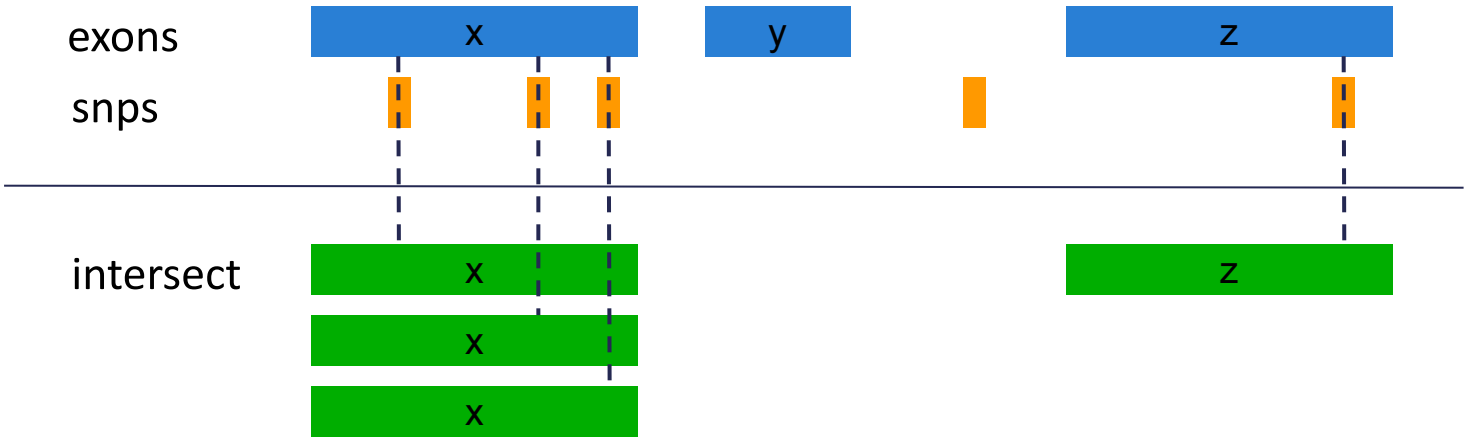
Which coding exon has the highest number of single nucleotide polymorphisms on chromosome 22?

exons.bed

Chrom	Start	End	Name	Score	Strand
chr22	16258185	16258303	uc002zlh.1_cds_1_0_chr22_16258186_r	0	-
chr22	16266928	16267095	uc002zlh.1_cds_2_0_chr22_16266929_r	0	-
chr22	16268136	16268181	uc002zlh.1_cds_3_0_chr22_16268137_r	0	-

snps.bed

Chrom	Start	End	Name	Score	Strand
chr22	16266919	16266920	rs757764551	0	+
chr22	16266920	16266920	rs375488594	0	+
chr22	16267014	16267015	rs544633418	0	+
chr22	16267029	16267030	rs563306354	0	+
chr22	16267081	16267082	rs568292779	0	+



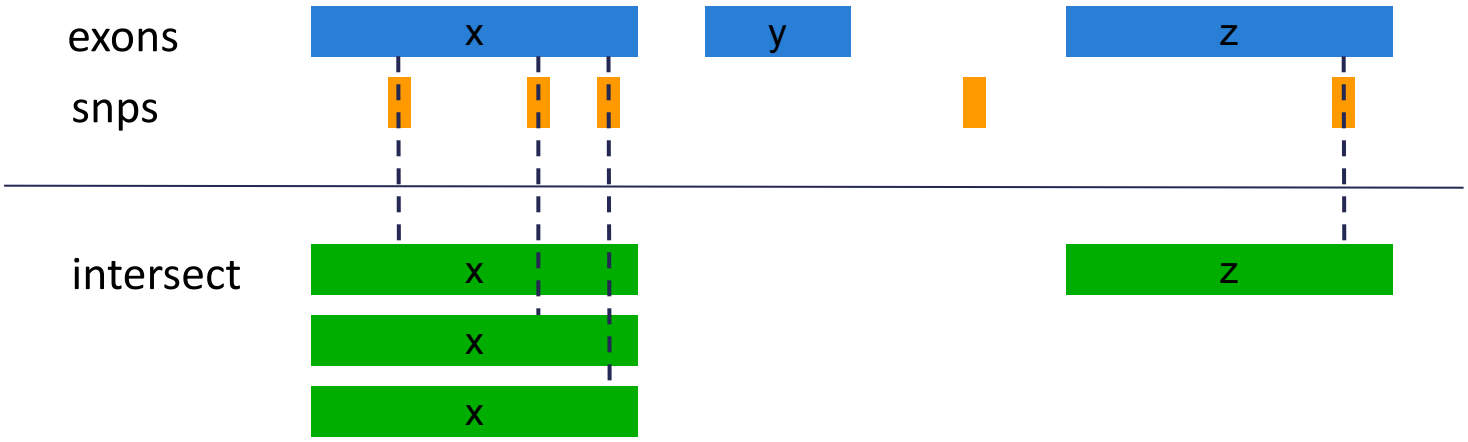
Which coding exon has the highest number of single nucleotide polymorphisms on chromosome 22?

exons.bed

Chrom	Start	End	Name	Score	Strand
chr22	16258185	16258303	uc002zlh.1_cds_1_0_chr22_16258186_r	0	-
chr22	16266928	16267095	uc002zlh.1_cds_2_0_chr22_16266929_r	0	-
chr22	16268136	16268181	uc002zlh.1_cds_3_0_chr22_16268137_r	0	-

snps.bed

Chrom	Start	End	Name	Score	Strand
chr22	16266919	16266920	rs757764551	0	+
chr22	16266920	16266920	rs375488594	0	+
chr22	16267014	16267015	rs544633418	0	+
chr22	16267029	16267030	rs563306354	0	+
chr22	16267081	16267082	rs568292779	0	+



exon	count
x	3
z	1
...	...

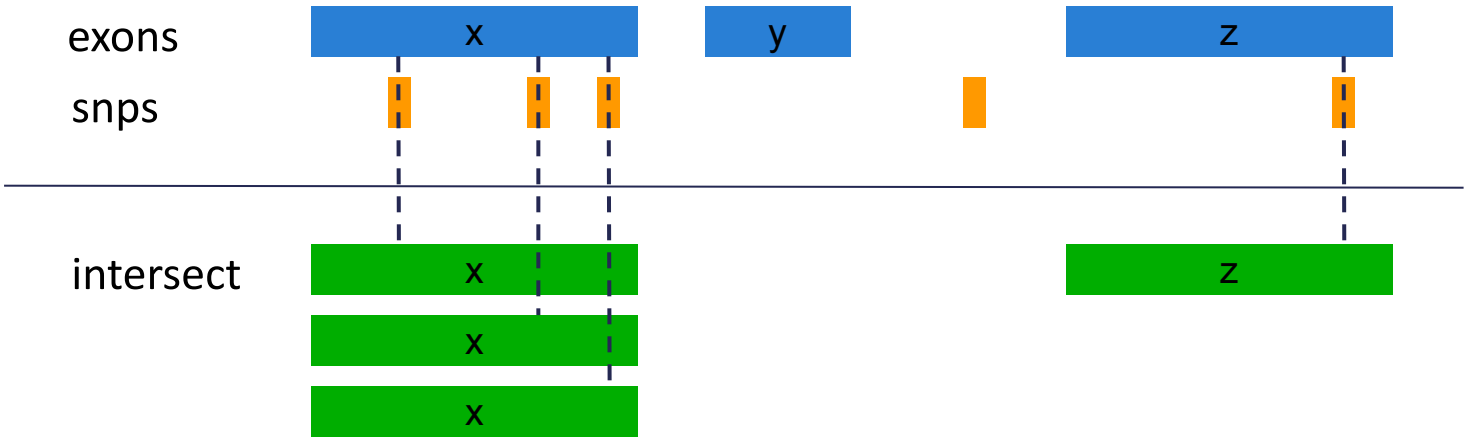
Which coding exon has the highest number of single nucleotide polymorphisms on chromosome 22?

exons.bed

Chrom	Start	End	Name	Score	Strand
chr22	16258185	16258303	uc002zlh.1_cds_1_0_chr22_16258186_r	0	-
chr22	16266928	16267095	uc002zlh.1_cds_2_0_chr22_16266929_r	0	-
chr22	16268136	16268181	uc002zlh.1_cds_3_0_chr22_16268137_r	0	-

snps.bed

Chrom	Start	End	Name	Score	Strand
chr22	16266919	16266920	rs757764551	0	+
chr22	16266920	16266920	rs375488594	0	+
chr22	16267014	16267015	rs544633418	0	+
chr22	16267029	16267030	rs563306354	0	+
chr22	16267081	16267082	rs568292779	0	+



exon	count
x	3
z	1
...	...

sort

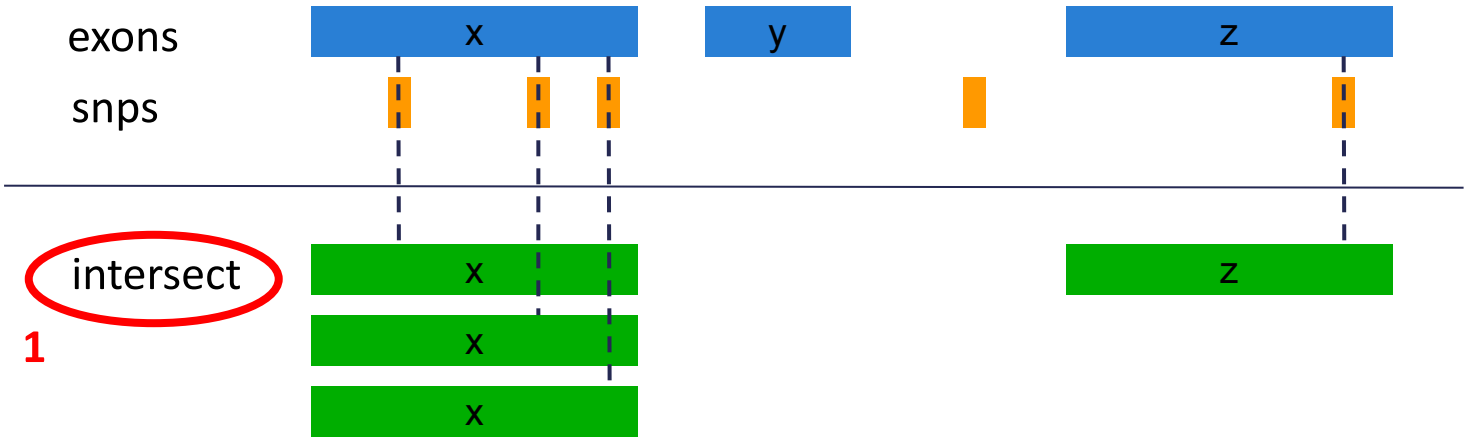
Which coding exon has the highest number of single nucleotide polymorphisms on chromosome 22?

exons.bed

Chrom	Start	End	Name	Score	Strand
chr22	16258185	16258303	uc002zlh.1_cds_1_0_chr22_16258186_r	0	-
chr22	16266928	16267095	uc002zlh.1_cds_2_0_chr22_16266929_r	0	-
chr22	16268136	16268181	uc002zlh.1_cds_3_0_chr22_16268137_r	0	-

snps.bed

Chrom	Start	End	Name	Score	Strand
chr22	16266919	16266920	rs757764551	0	+
chr22	16266920	16266920	rs375488594	0	+
chr22	16267014	16267015	rs544633418	0	+
chr22	16267029	16267030	rs563306354	0	+
chr22	16267081	16267082	rs568292779	0	+



2

exon	count
x	3
z	1
...	...

sort

3

Tools – Hands-on

Which coding exon has the highest number of single nucleotide polymorphisms on chromosome 22?

1. Get the data in a new history
2. Intersect exons with SNPs
3. Count the number of SNPs per exon
4. Sort exons by SNP count
5. Select top five
6. Recover exon info
7. Display data in UCSC genome browser

Which coding exon has the highest number of single nucleotide polymorphisms on chromosome 22?

1. Get the data in a new history
2. Intersect exons with SNPs *Bedtools -> Intersect intervals*
3. Count the number of SNPs per exon
4. Sort exons by SNP count
5. Select top five
6. Recover exon info
7. Display data in UCSC genome browser

Q : For the first 3 exons in your file, what is the number of SNPs that fall into that exon?

Tools – Hands-on

Which coding exon has the highest number of single nucleotide polymorphisms on chromosome 22?

1. Get the data in a new history
2. Intersect exons with SNPs *Bedtools -> Intersect intervals*
3. Count the number of SNPs per exon *Join, Subtract, and Group -> Group*
4. Sort exons by SNP count
5. Select top five
6. Recover exon info
7. Display data in UCSC genome browser

Q : How many exons are there in total in your file?

Tools – Hands-on

Which coding exon has the highest number of single nucleotide polymorphisms on chromosome 22?

1. Get the data in a new history
2. Intersect exons with SNPs *Bedtools -> Intersect intervals*
3. Count the number of SNPs per exon *Join, Subtract, and Group -> Group*
4. Sort exons by SNP count *Filter and Sort -> Sort*
5. Select top five
6. Recover exon info
7. Display data in UCSC genome browser

Q : Which exon has the highest number of SNPs in your file?

Tools – Hands-on

Which coding exon has the highest number of single nucleotide polymorphisms on chromosome 22?

1. Get the data in a new history
2. Intersect exons with SNPs *Bedtools -> Intersect intervals*
3. Count the number of SNPs per exon *Join, Subtract, and Group -> Group*
4. Sort exons by SNP count *Filter and Sort -> Sort*
5. Select top five *Text Manipulation -> Select First*
6. Recover exon info
7. Display data in UCSC genome browser

Which coding exon has the highest number of single nucleotide polymorphisms on chromosome 22?

1. Get the data in a new history
2. Intersect exons with SNPs *Bedtools -> Intersect intervals*
3. Count the number of SNPs per exon *Join, Subtract, and Group -> Group*
4. Sort exons by SNP count *Filter and Sort -> Sort*
5. Select top five *Text Manipulation -> Select First*
6. Recover exon info *Join, Subtract and Group -> Compare two Datasets*
7. Display data in UCSC genome browser

BED file

Chrom	Start	End	Name	Score	Strand
chr22	16258185	16258303	uc002zlh.1_cds_1_0_chr22_16258186_r	0	-
chr22	16266928	16267095	uc002zlh.1_cds_2_0_chr22_16266929_r	0	-

Cleanup

DATASET

Dataset

Delete a dataset

Dataset

The dataset isn't really deleted. It's in the Trash

Galaxy / ABiMS | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 0%

Tools

search tools

Get Data
Send Data
Collection Operations

COMMON TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Statistics
Graph/Display Data
Fasta Fastq Manipulation

COMMON NGS TOOLS
NGS:Samtools
NGS:Mapping
NGS:Bedtools
NGS:Picard Tools

SEARCHING TOOLS
Diamond

javascript:void(0);

Welcome to galaxy3.sb-roscoff.fr

Information
For any question or request for tools or account, send an email at support.abims@sb-roscoff.fr

ABiMS
Analyses and Bioinformatics for Marine Science
Station Biologique Roscoff

Changelog

Tutorials

Galaxy is an open, web-based platform for data intensive biomedical research. 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 NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

History

search datasets

eba 2016 sartools
42 shown, hide deleted

1.59 MB

This dataset has been deleted
Undelete it
Permanently remove it from disk

62: SARTools DESeq2 R objects (.RData)

61: SARTools DESeq2 R log

60: SARTools DESeq2 figures

59: SARTools DESeq2 tables

58: SARTools DESeq2 report

57: SARTools edgeR R objects (.RData)

56: SARTools edgeR

Dataset

"Empty Trash" : to free up disk space

Galaxy / ABiMS | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 0%

Tools

search tools

Get Data
Send Data
Collection Operations

COMMON TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Statistics
Graph/Display Data
Fasta Fastq Manipulation

COMMON NGS TOOLS
NGS:Samtools
NGS:Mapping
NGS:Bedtools
NGS:Picard Tools

SEARCHING TOOLS
Diamond

Welcome to galaxy3.sb-roscoff.fr

Information
For any question or request for tools or account, send an email at support.abims@sb-roscoff.fr

ABiMS
Analyses and Bioinformatics for Marine Science
Station Biologique Roscoff

Changelog
Tutorials

Galaxy is an open, web-based platform for data intensive biomedical research. 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 NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

History

HISTORY LISTS
Saved Histories
Histories Shared with Me
HISTORY ACTIONS
Create New
Copy History
Share or Publish
Show Structure
Extract Workflow
Delete
Delete Permanently
DATASET ACTIONS
Copy Datasets
Dataset Security
Resume Paused Jobs
Collapse Expanded Datasets
Unhide Hidden Datasets
Delete Hidden Datasets
Purge Deleted Datasets
DOWNLOADS
Export Tool Citations
Export History to File
OTHER ACTIONS
Import from File

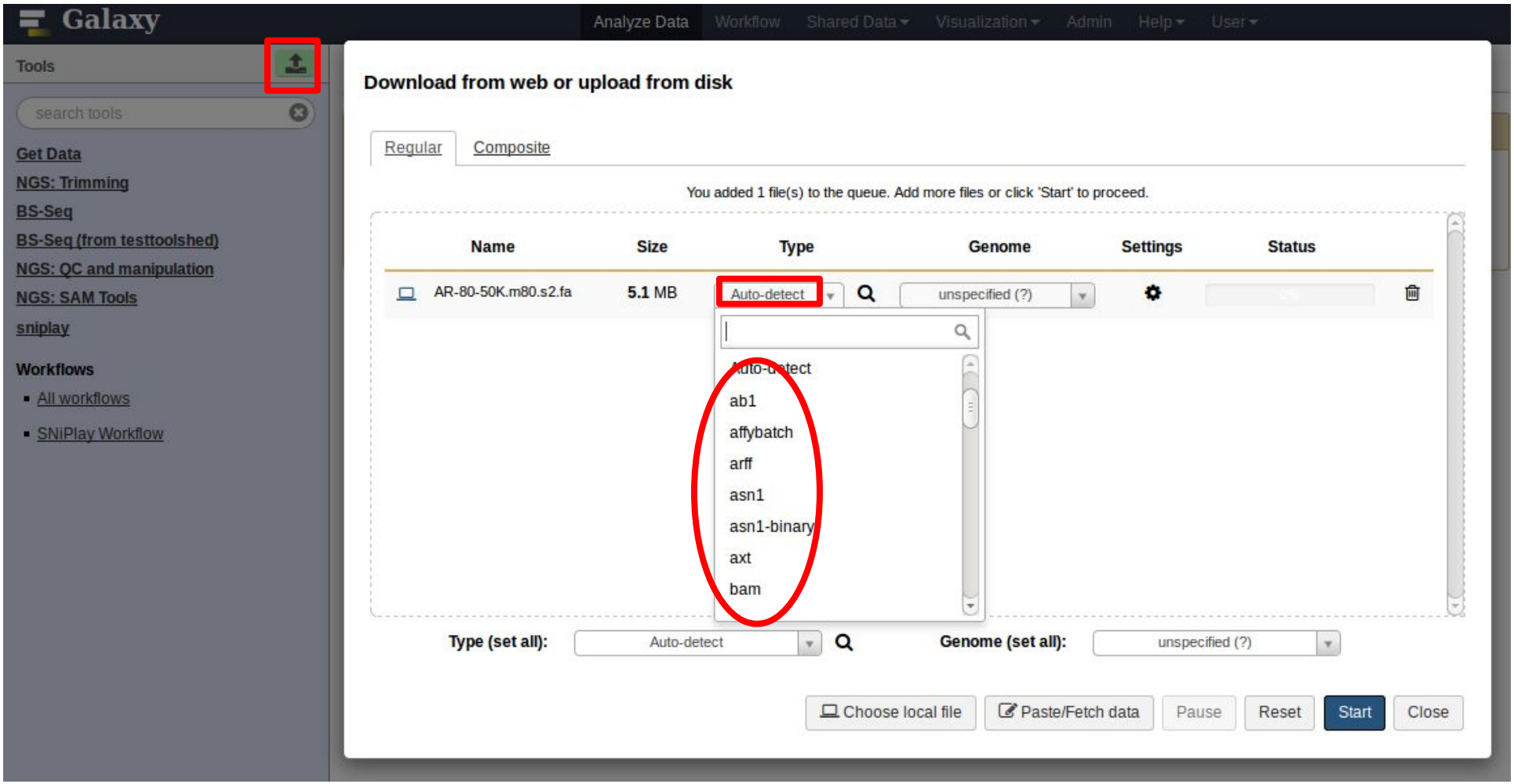
galaxy3.sb-roscoff.fr/history/purge_deleted_datasets

Datatypes

DATASET

Dataset - Datatypes

- Every Galaxy dataset is associated with a datatype.
- Datatype can be detected or user specified.



Dataset - Datatypes

Common text formats:

- ***txt***: plain text ('.txt')
- ***tabular***: tab delimited ('.tab', '.txt', etc.)

```
                  wt_37_2     wt_37_3     wt_37_1
TR24|c0_g1_i1     90.00       67.00       85.00
TR2779|c0_g1_i1   186.00      137.00      217.00
TR127|c1_g1_i1    9.00        23.00       16.00
```

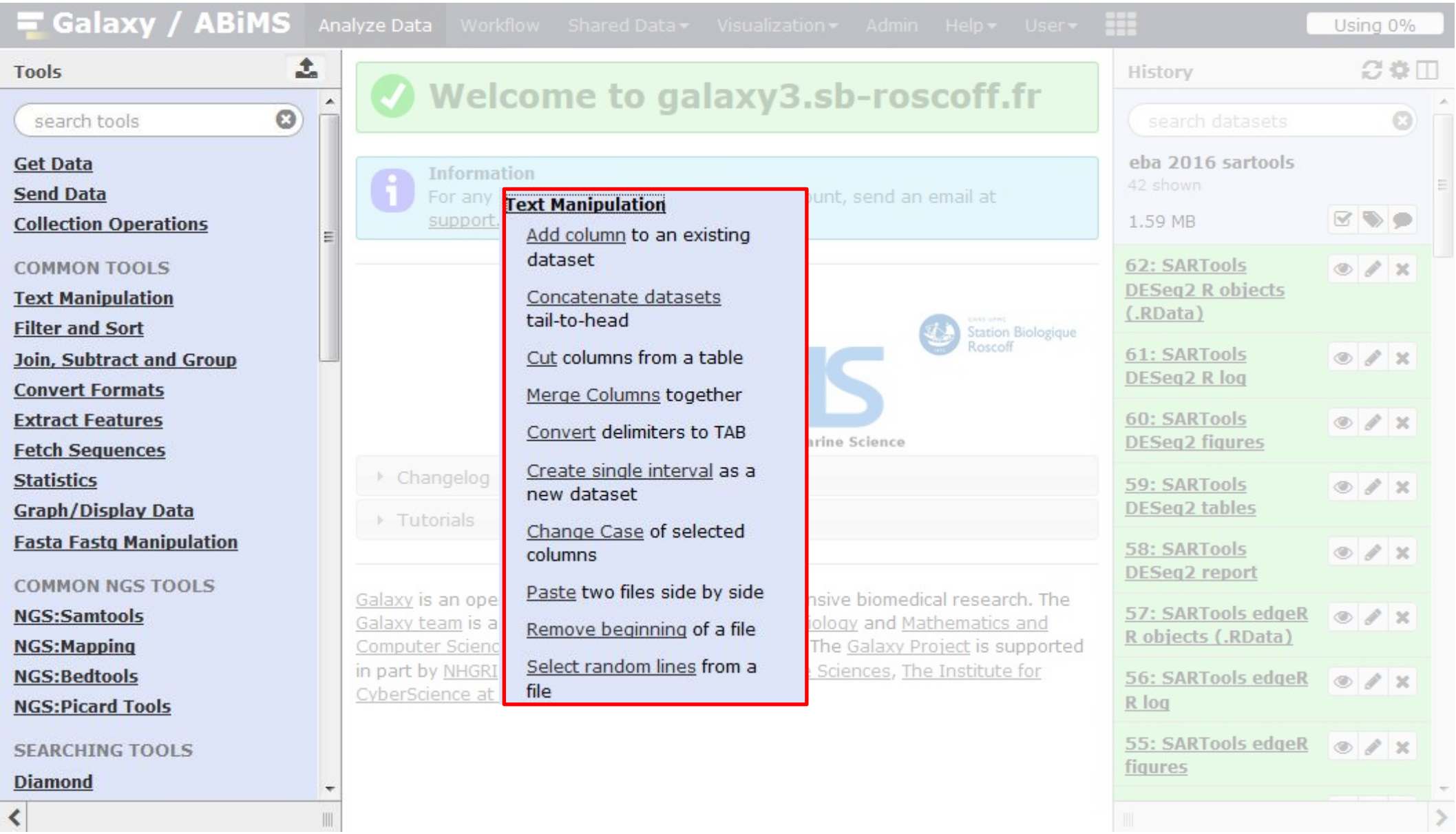
- ***csv***: comma-separated values ('.csv')

```
Year,Make,Model
1997,Ford,E350
2000,Mercury,Couga
r
```

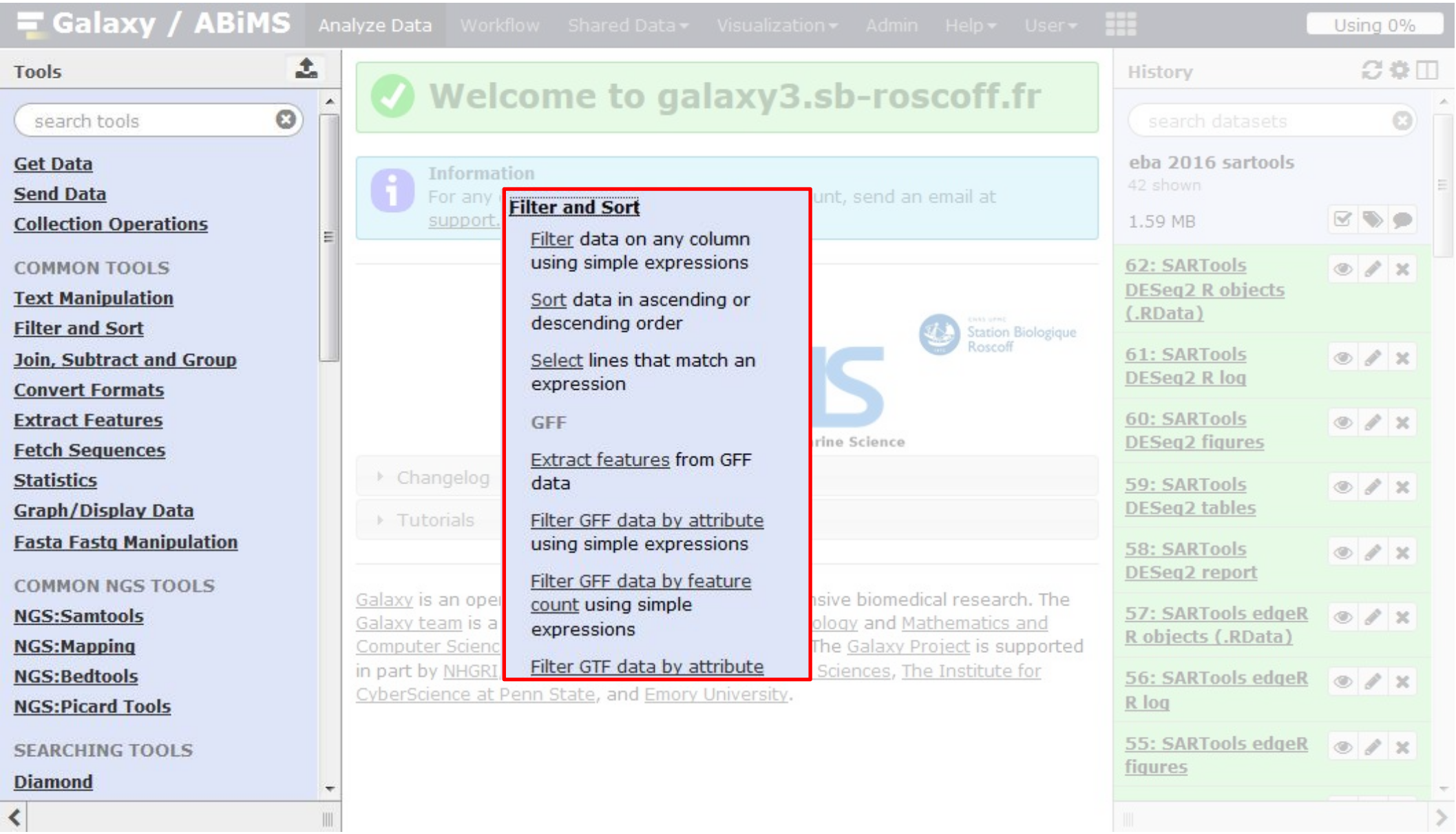
- ***html***: standard language for web pages

```
<!DOCTYPE html>
<html>
  <head>
    <title>This is a
title</title>
  </head>
  <body>
    <p>Hello world!</p>
  </body>
</html>
```

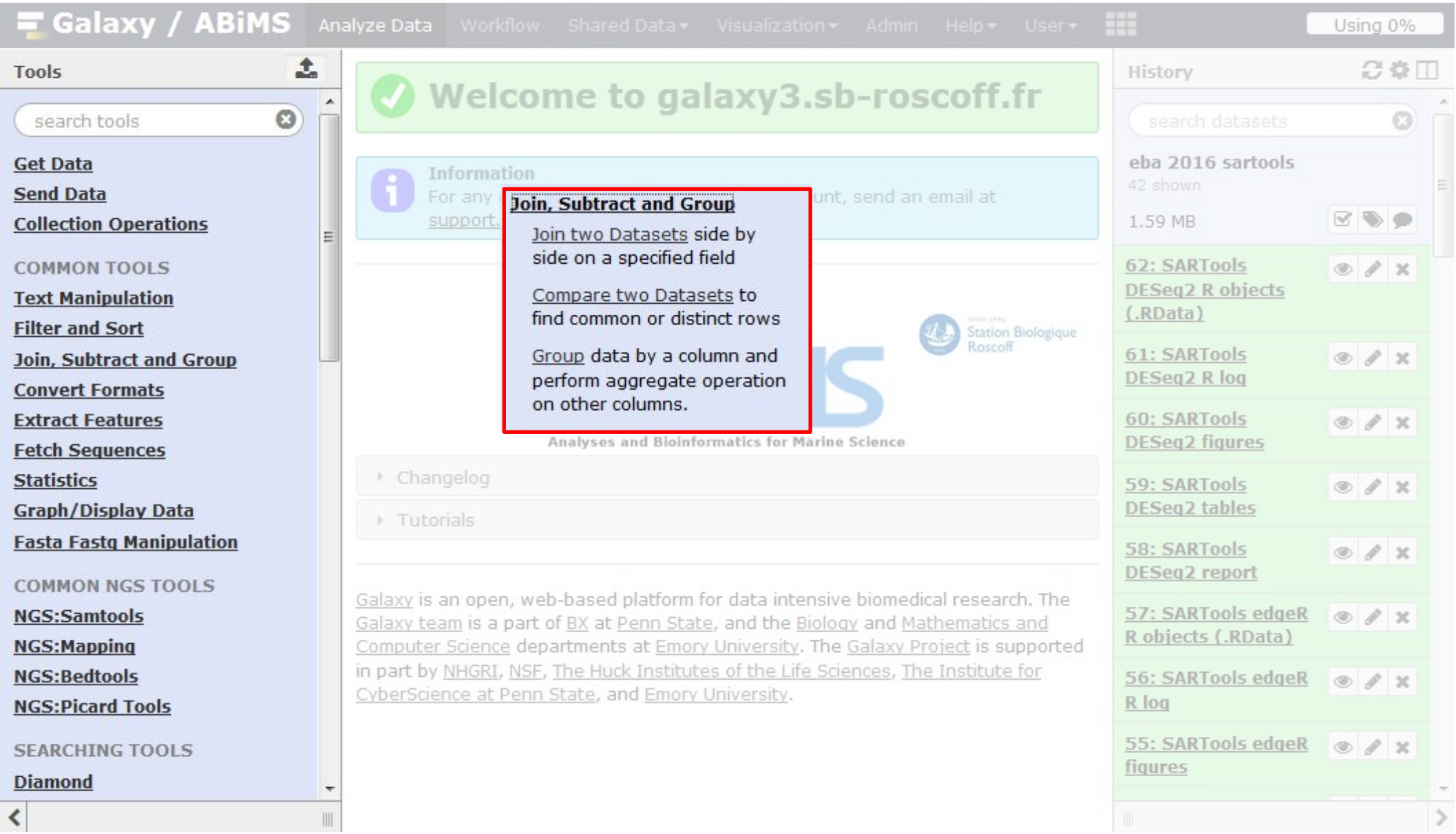
Dataset - Datatypes



Dataset - Datatypes



Dataset - Datatypes



Dataset - Datatypes

Common binary formats:

- ***data***: generic binary format
- ***zip, tar***: archives
- ***pdf, png, jpg, bmp, tiff, gif***: images
- ***rdata***: statistical computing program R
- ***bam*, wig, bigwig**: sequence alignment

Dataset - Datatypes

Sequence file formats:

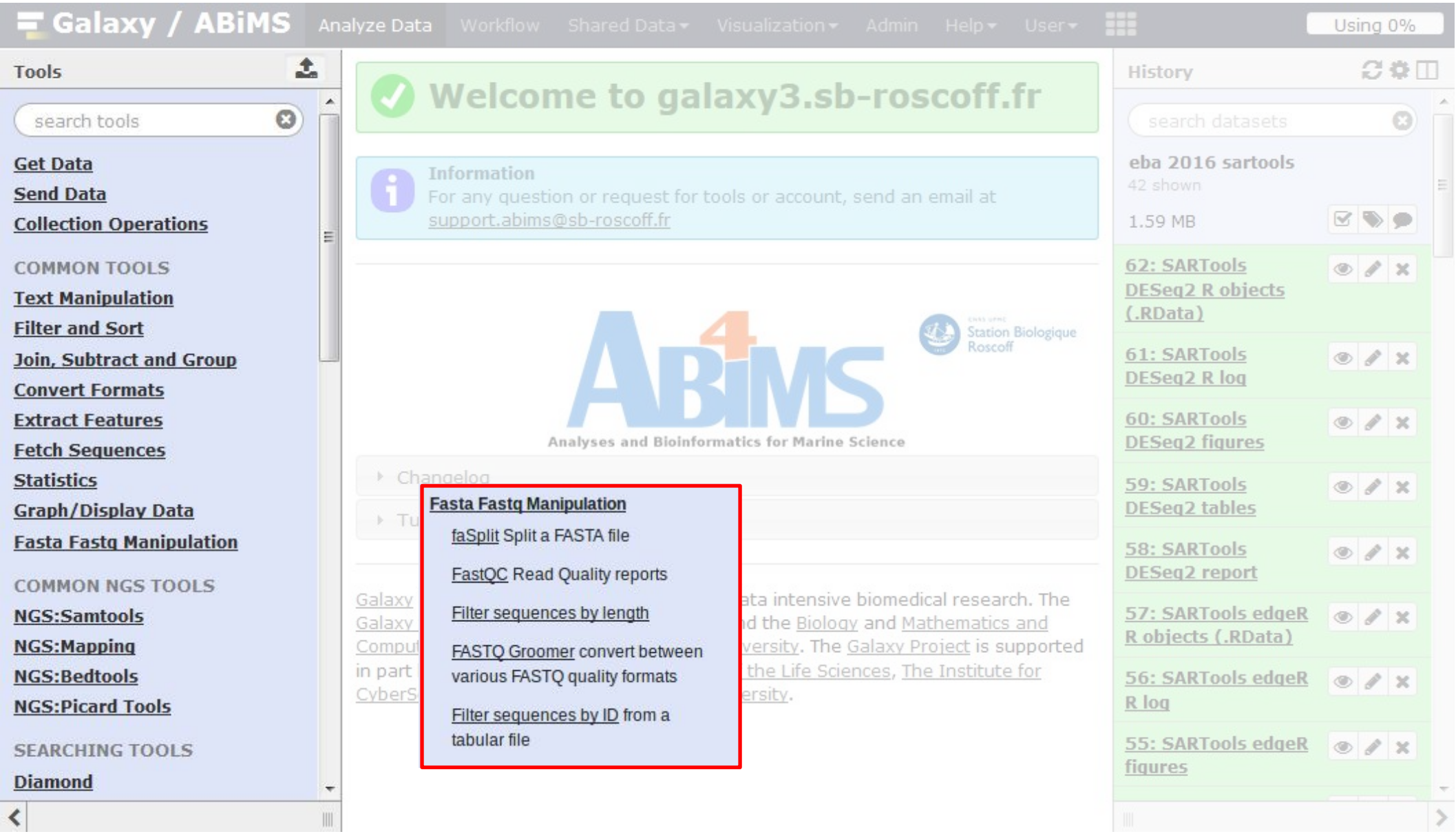
- ***fasta***: a single-line description with '>', followed by lines of sequence data ('.fasta', '.fas')

```
>sequence1
atgcgtttgcgtgcatgcgtttgcgtgcatgcgtttgcgtgcatgcgtttgcgtg
c
atgcgtttgcgtgc
>sequence2
tttcgtgcgtatagtttcgtgcgtatagtttcgtgcgtatagtttcgtgcgtata
g
tggcgcggt
```

- ***fastq***: sequence + quality score ('.fastq', '.fq')

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTT
T
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC6
5
@SEQ_ID2
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTT
T
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC6
```

Dataset - Datatypes



Dataset - Datatypes

Sequence file formats:

- ***gff3, bed, genbank***: sequence + annotations

bed

```
track name=pairedReads description="Clone Paired Reads"
useScore=1
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512
chr22 2000 6000 cloneB 900 - 2000 6000 0 2 433,399, 0,3601
```

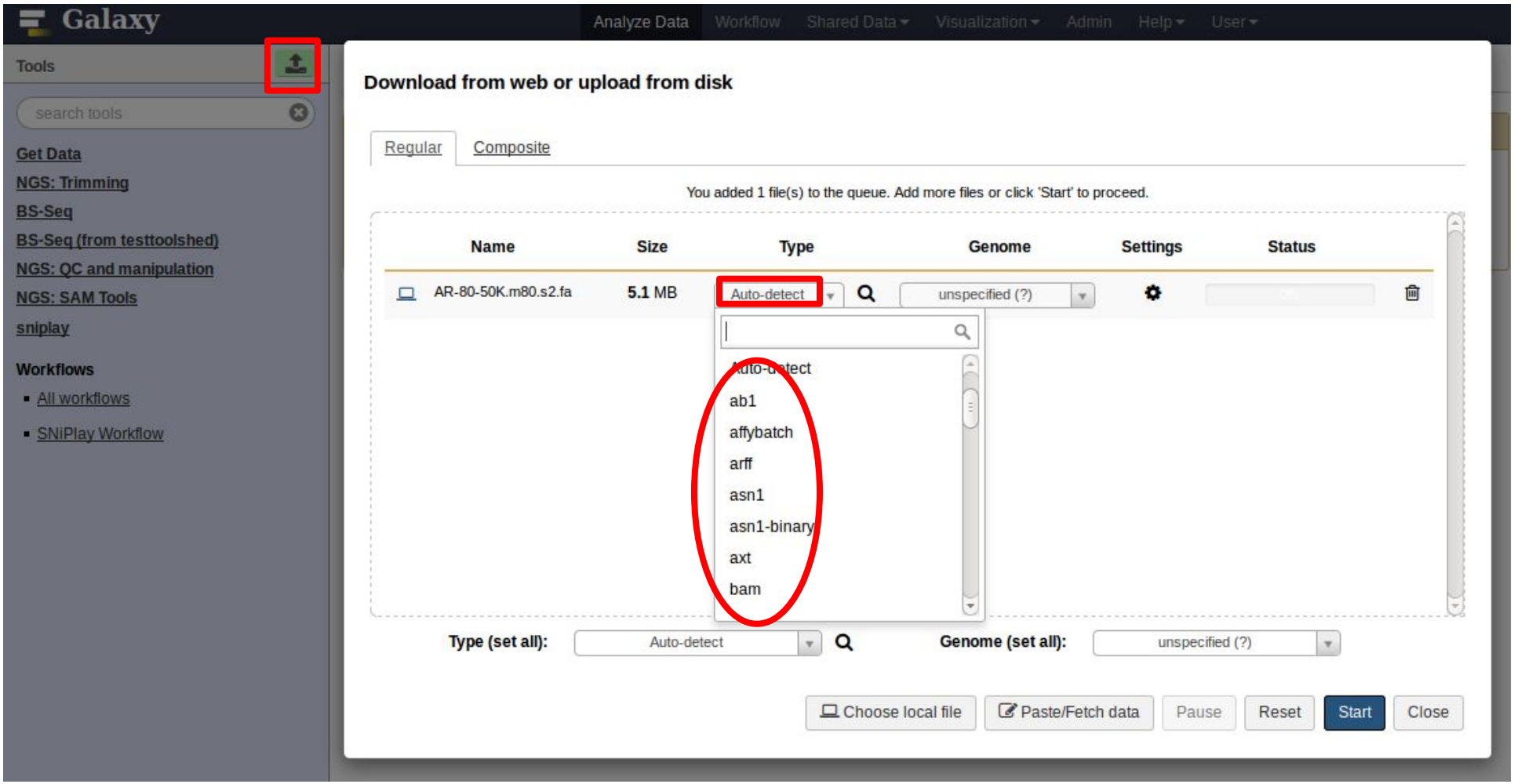
gff3

```
##gff-version 3
ctg123  .  exon  1300  1500  .  +  .
ID=exon00001
ctg123  .  exon  1050  1500  .  +  .
ID=exon00002
ctg123  .  exon  3000  3902  .  +  .
ID=exon00003
##FASTA
>ctg123
cttctgggcgtacccgattctcggagaacttgccgcaccattccgcctt
g
tgttcattgctgcctgcatgttcattgtctacctcggctacgtgtggct
g
…
```

https://genome.ucsc.edu/FAQ/FAQformat.html#format3

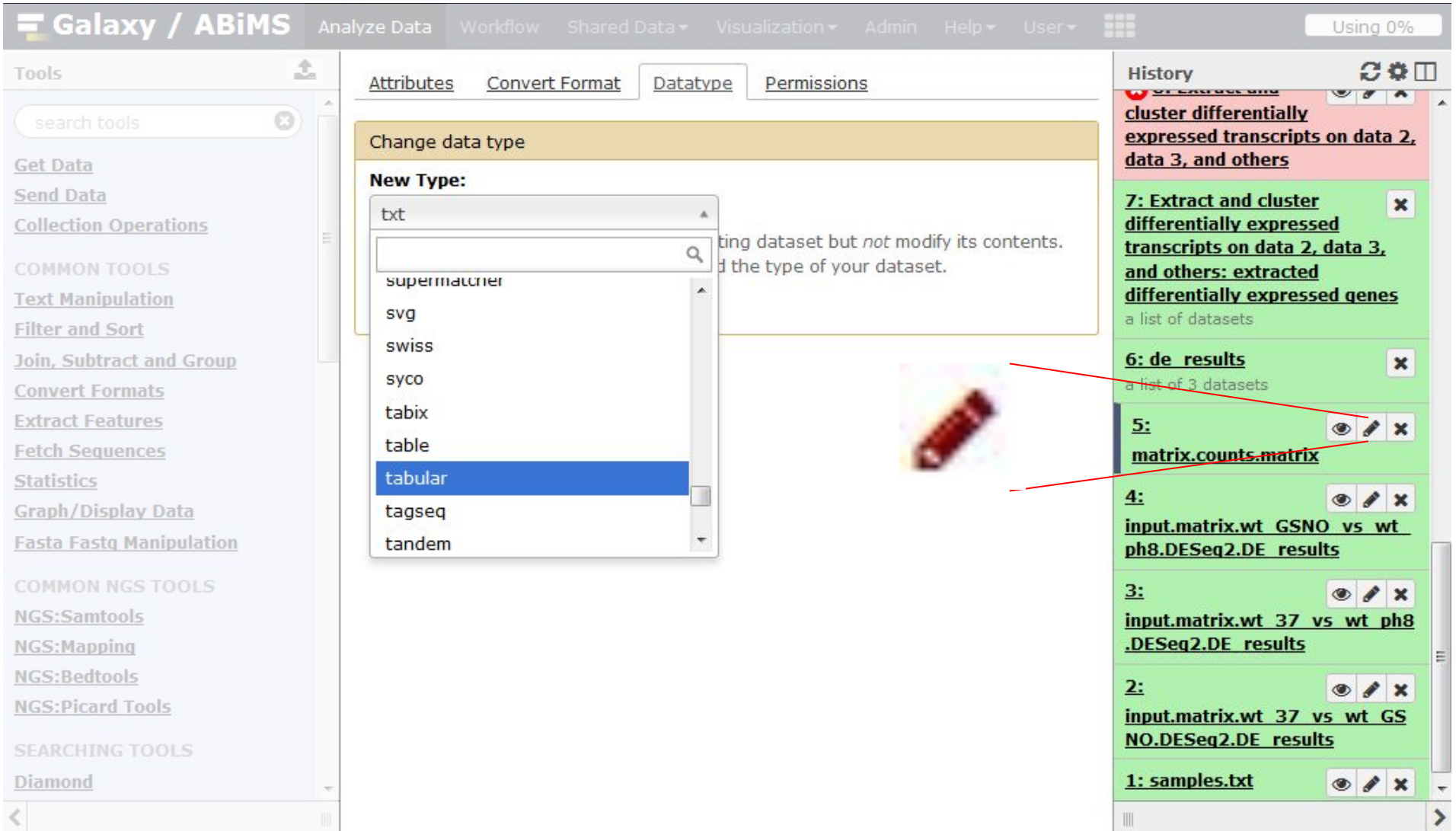
Dataset - Datatypes

- Every Galaxy dataset is associated with a datatype.
- Datatype can be detected or user specified.



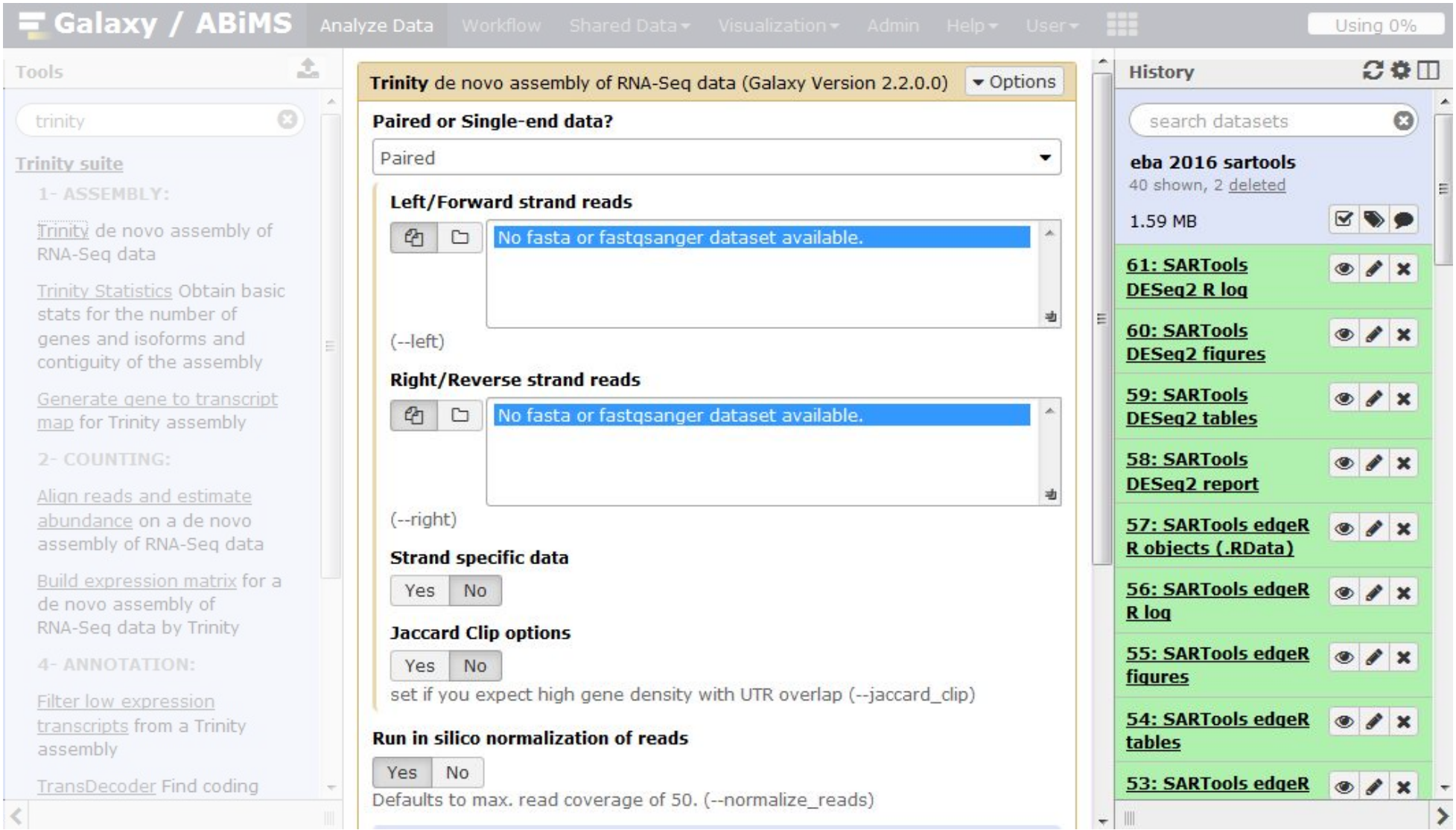
Dataset - Datatypes

- Every Galaxy dataset is associated with a datatype.
- Datatype can be detected or user specified.



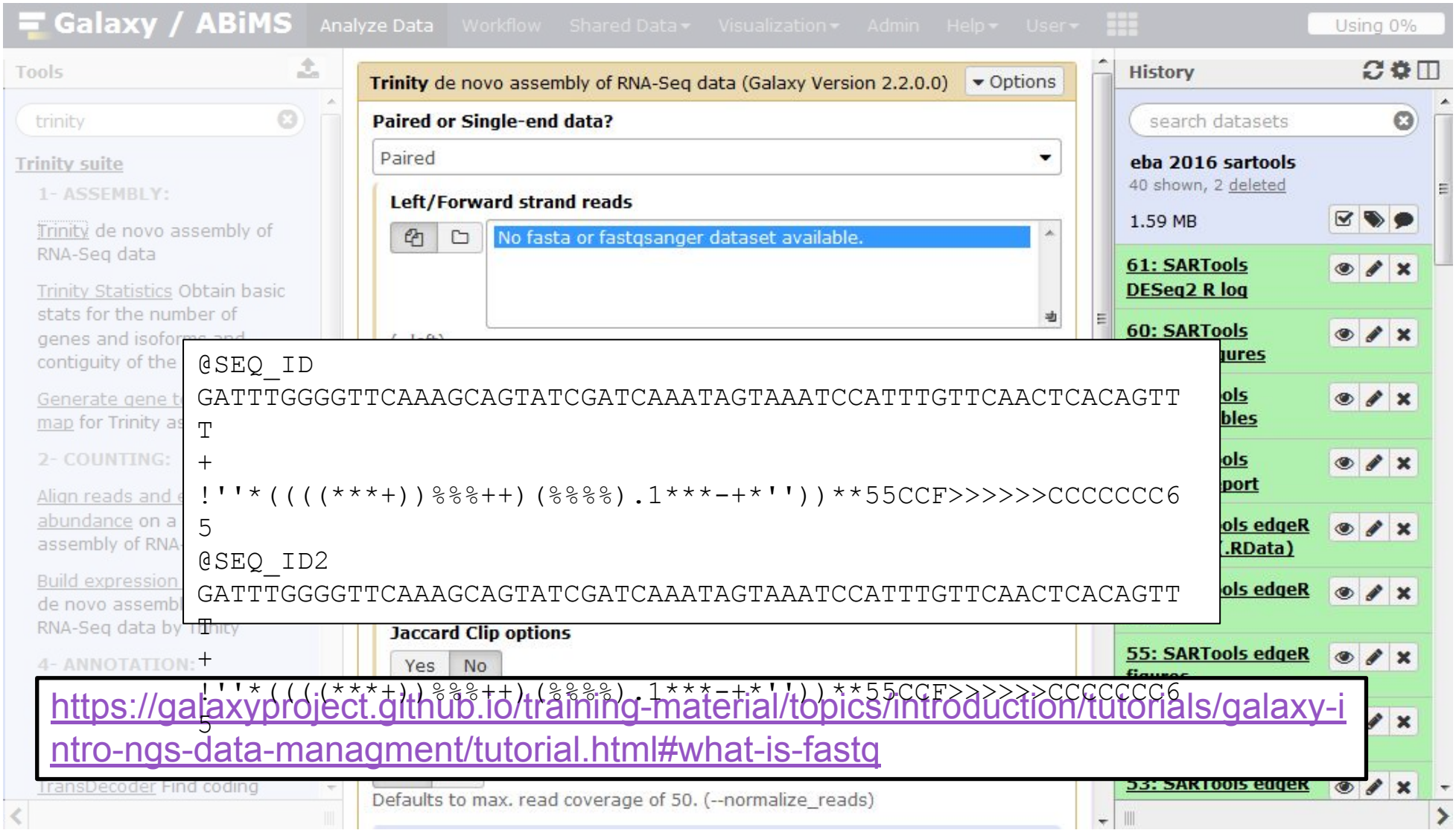
Dataset - Datatypes

- Many tools will only accept input datasets with the appropriate datatype assigned.



Dataset - Datatypes

- Many tools will only accept input datasets with the appropriate datatype assigned.



https://galaxyproject.github.io/training-material/topics/introduction/tutorials/galaxy-intro-ngs-data-managment/tutorial.html#what-is-fastq

Hands-on

TOOLS (sequence files manipulation)

Tools (sequence files manipulation) – Hands-on

Part 1:

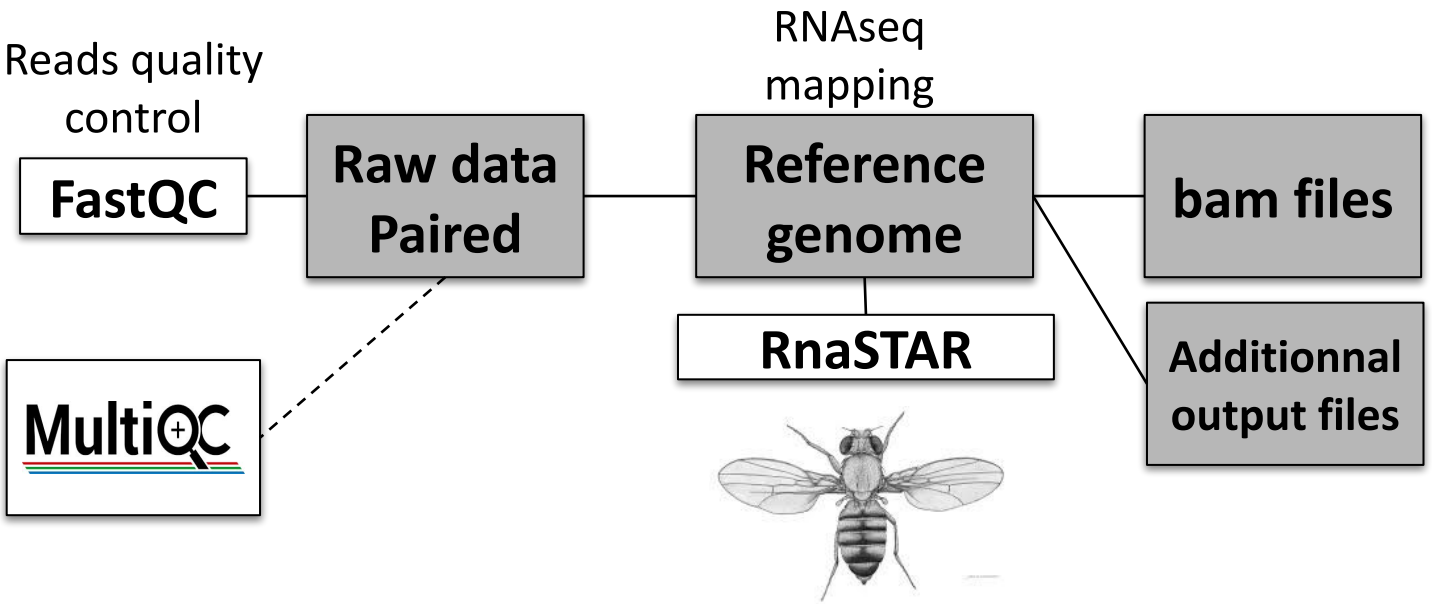
You have sequencing data from your favorite species. You want to check the quality of your sequences and to map on the reference genome.

Case of Drosophila melanogaster (dm)

Part 1:

You have sequencing data from your favorite species. You want to check the quality of your sequences and to map on the reference genome.

Case of Drosophila melanogaster (dm)



Tools (sequence files manipulation) – Hands-on

Part 1:

You have sequencing data from your favorite species. You want to check the quality of your sequences and to map on the reference genome.

Case of Drosophila melanogaster (dm)

1. Look into the description of the tool FastQC, what is its purpose ?
2. Does it takes compressed files ? ".gz"
3. Which encoding of the file (in FastQC results, basic statistics) ? => edit if needed the datatype of your reads files
4. Look at mapping tools, map your reads on genome with RNA STAR.
5. What is the default parameter of the option "Maximum ratio of mismatches to mapped length" in RNA STAR ?
6. How to map on another reference genome or assembly release ?

Part 2:

You have new sequences that you want to compare with the gene and protein databases from your favorite species (BLAST).

Case of Drosophila melanogaster (dm)

1. Create new history
2. Import CDS and peptide sequences databases
3. Import query sequences
4. Make BLAST databases
5. Run BLAST against the CDS database
6. Run BLAST against the protein database

Part 2:

You have new sequences that you want to compare with the gene and protein databases from your favorite species (BLAST).

Case of Drosophila melanogaster (dm)

1. Create new history
2. Import CDS and peptide sequences databases
3. Import query sequences
4. Make BLAST databases *Blast --> NCBI BLAST+ makeblastdb*
5. Run BLAST against the CDS database
6. Run BLAST against the protein database

Tools (sequence files manipulation) – Hands-on

Part 2:

You have new sequences that you want to compare with the gene and protein databases from your favorite species (BLAST).

Case of Drosophila melanogaster (dm)

1. Create new history
2. Import CDS and peptide sequences databases
3. Import query sequences
4. Make BLAST databases *Blast --> NCBI BLAST+ makeblastdb*
5. Run BLAST against the CDS database *Blast --> NCBI BLAST+ blastn*
6. Run BLAST against the protein database

Tools (sequence files manipulation) – Hands-on

Part 2:

You have new sequences that you want to compare with the gene and protein databases from your favorite species (BLAST).

Case of Drosophila melanogaster (dm)

1. Create new history
2. Import CDS and peptide sequences databases
3. Import query sequences
4. Make BLAST databases *Blast --> NCBI BLAST+ makeblastdb*
5. Run BLAST against the CDS database *Blast --> NCBI BLAST+ blastn*
6. Run BLAST against the protein database *Blast --> NCBI BLAST+ blastx*

DATASET COLLECTION

Dataset collection

- Problematic: you have a large numbers of datasets to send through the same analysis

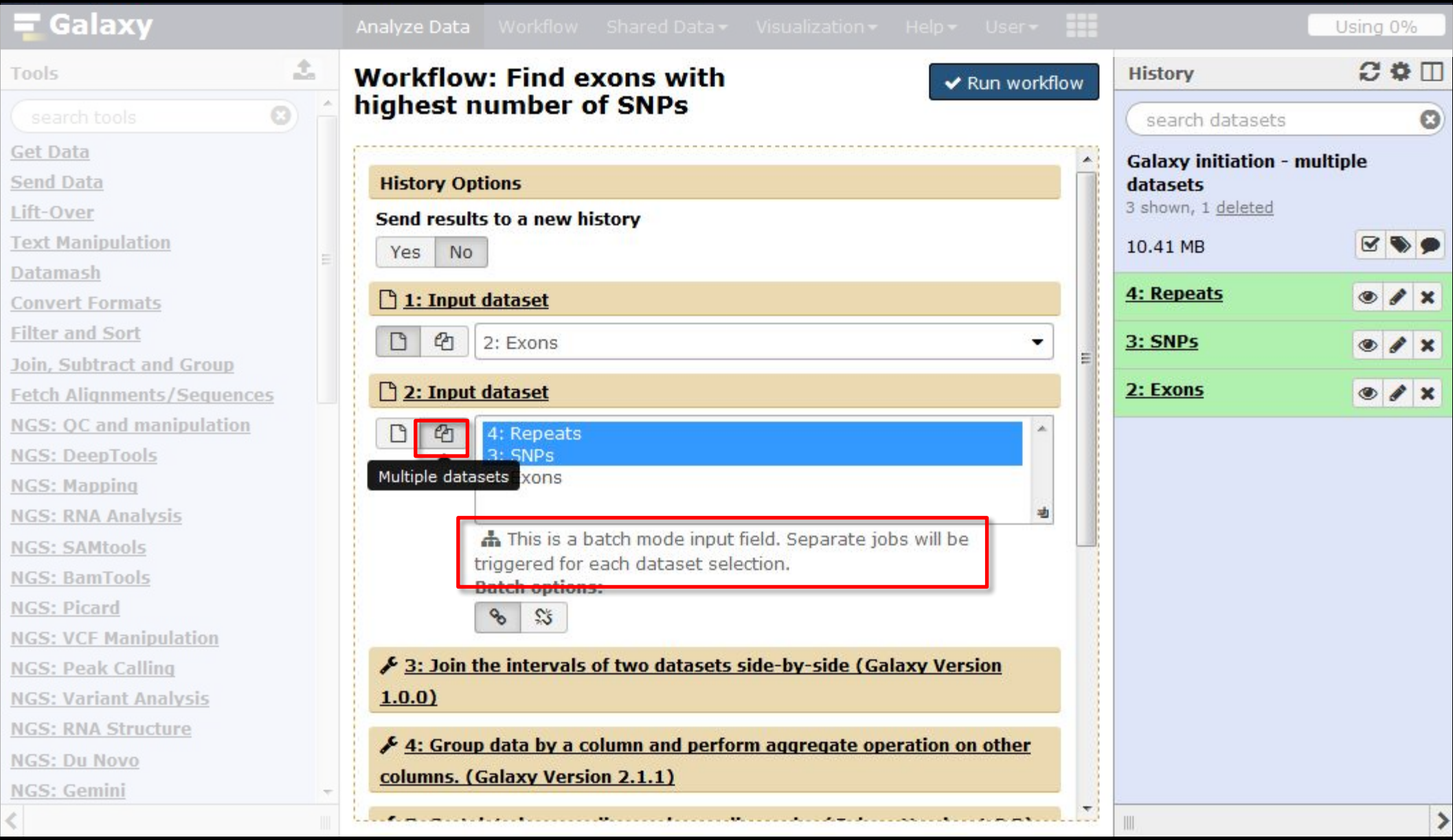
Dataset collection

- Problematic: you have a large numbers of datasets to send through the same analysis

- Solution 1: select multiple datasets as input

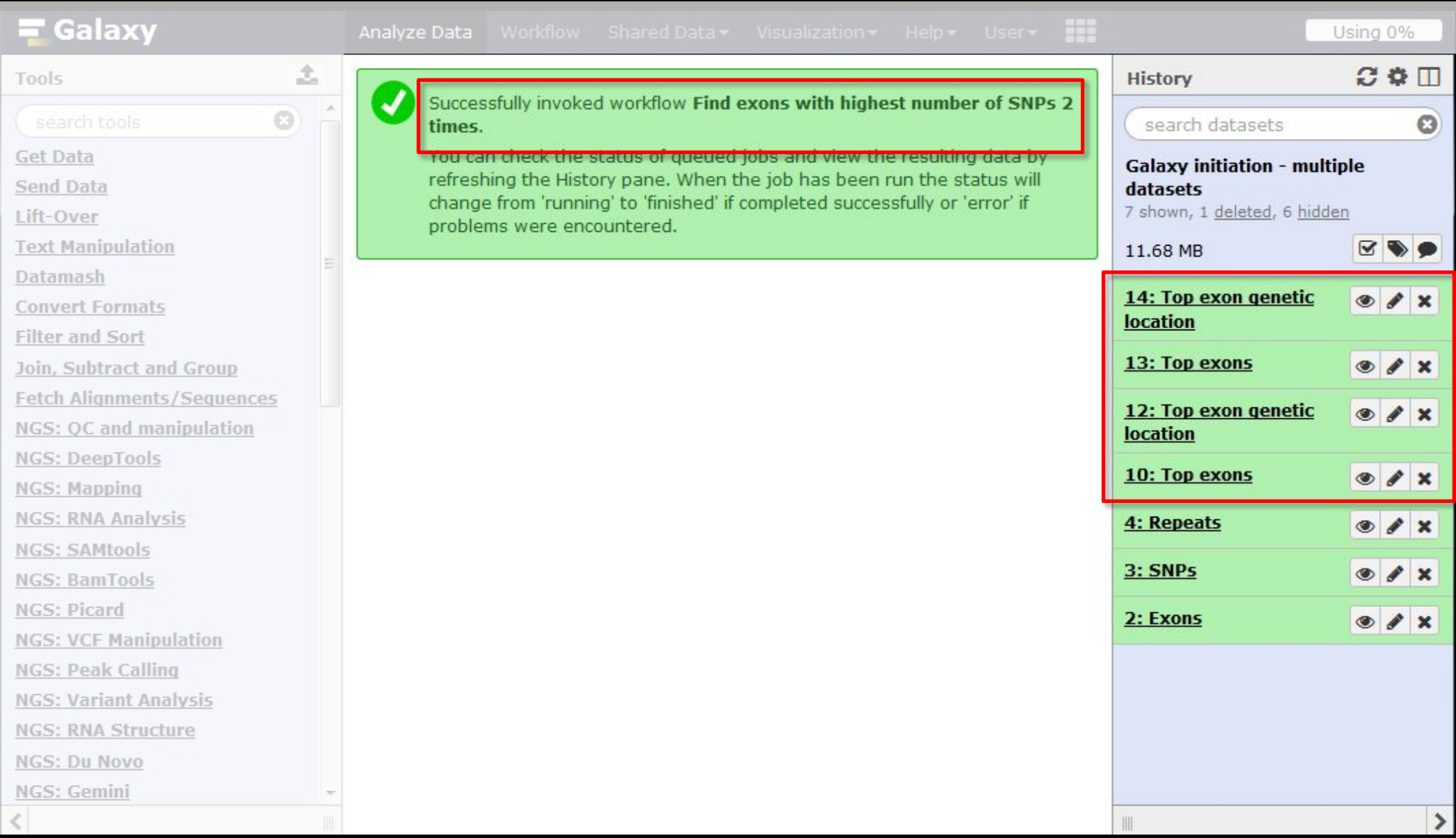
Dataset collection

Select multiple datasets as input



Dataset collection

Select multiple datasets as input



Dataset collection

- Problematic: you have a large numbers of datasets to send through the same analysis

- Solution 1: select multiple datasets as input
- Solution 2: create a dataset collection

Dataset collection

- Problematic: you have a large numbers of datasets to send through the same analysis

- Solution 1: select multiple datasets as input
- Solution 2: create a dataset collection
  - Dataset list: set of files of the same type
  - Dataset pairs: pairs of read files (forward, reverse)
  - List of dataset pairs

Dataset collection

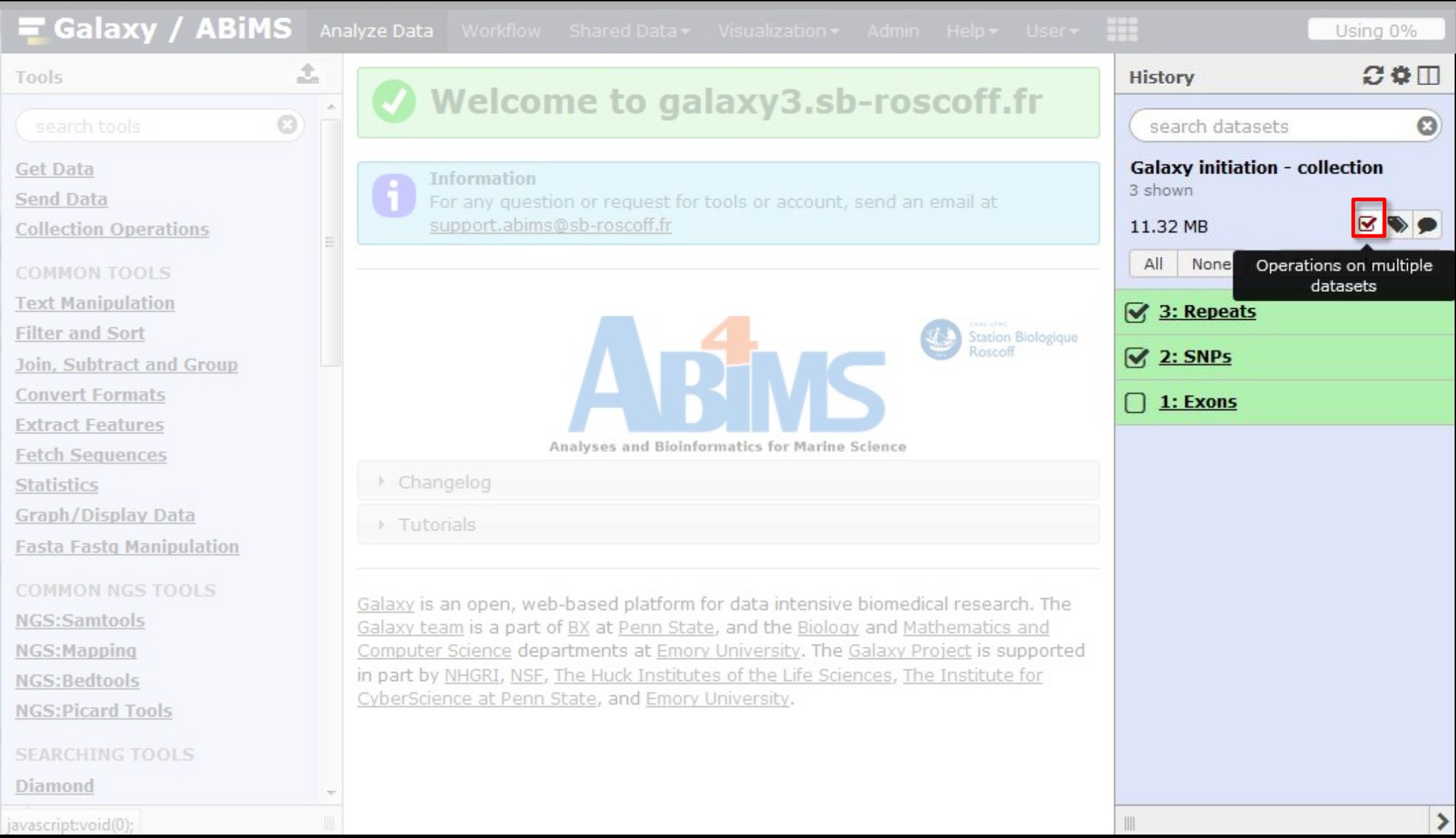
- Problematic: you have a large numbers of datasets to send through the same analysis

- Solution 1: select multiple datasets as input
- Solution 2: create a dataset collection
  - Dataset list: set of files of the same type
  - Dataset pairs: pairs of read files (forward, reverse)
  - List of dataset pairs

- Galaxy runs the tool automatically on each dataset in the collection using the same settings

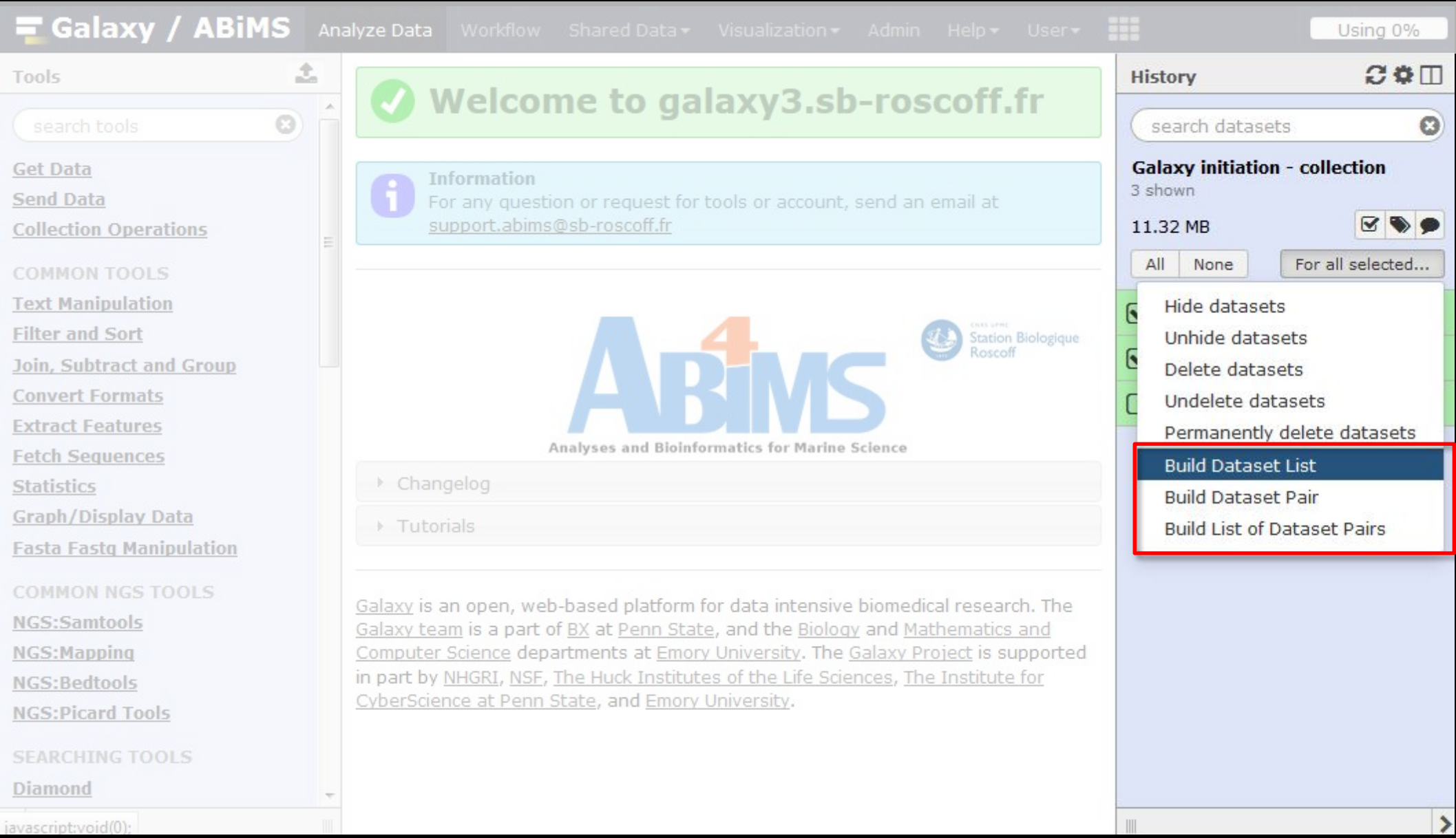
Dataset collection

Create a dataset collection



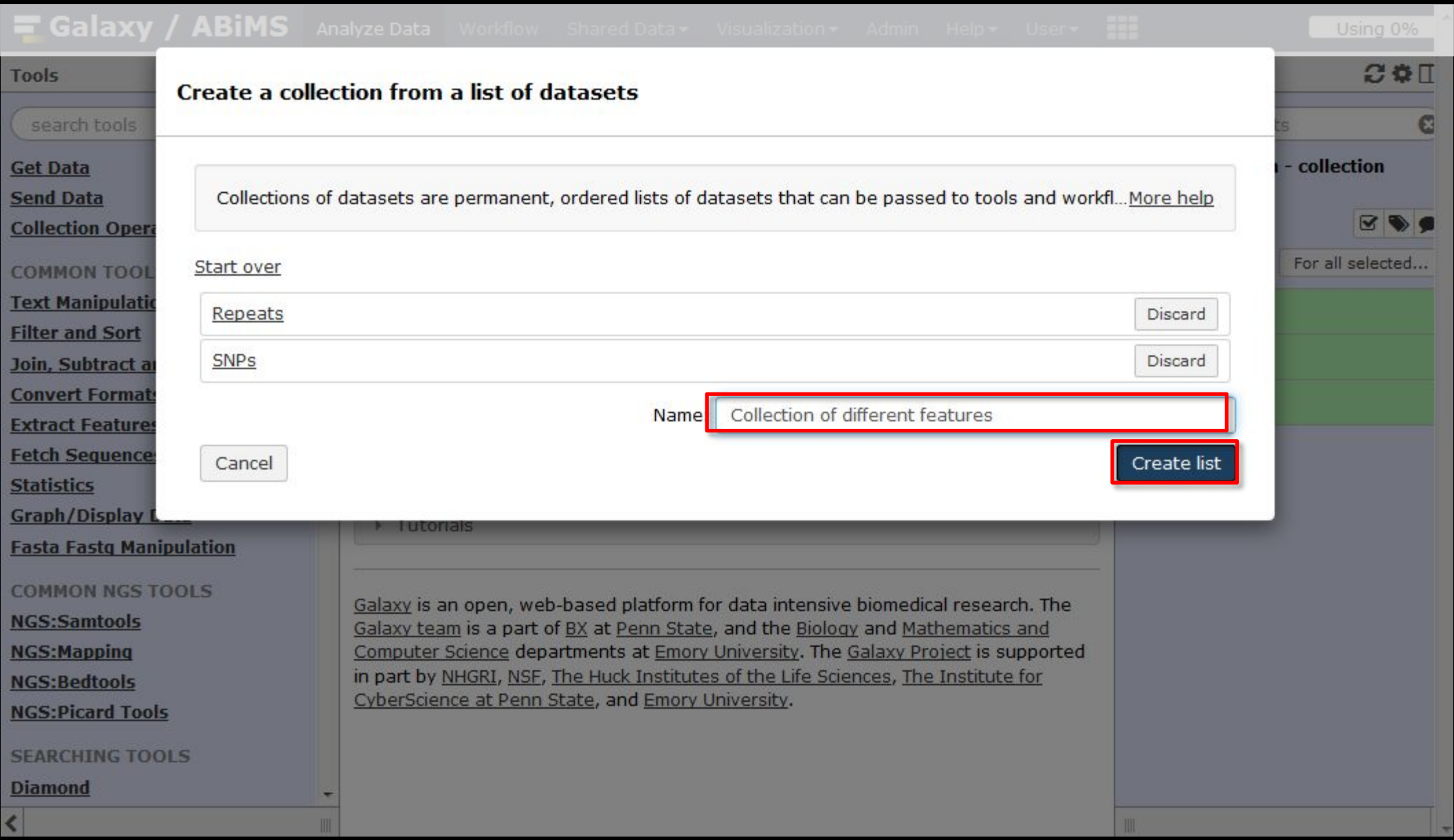
Dataset collection

Create a dataset collection



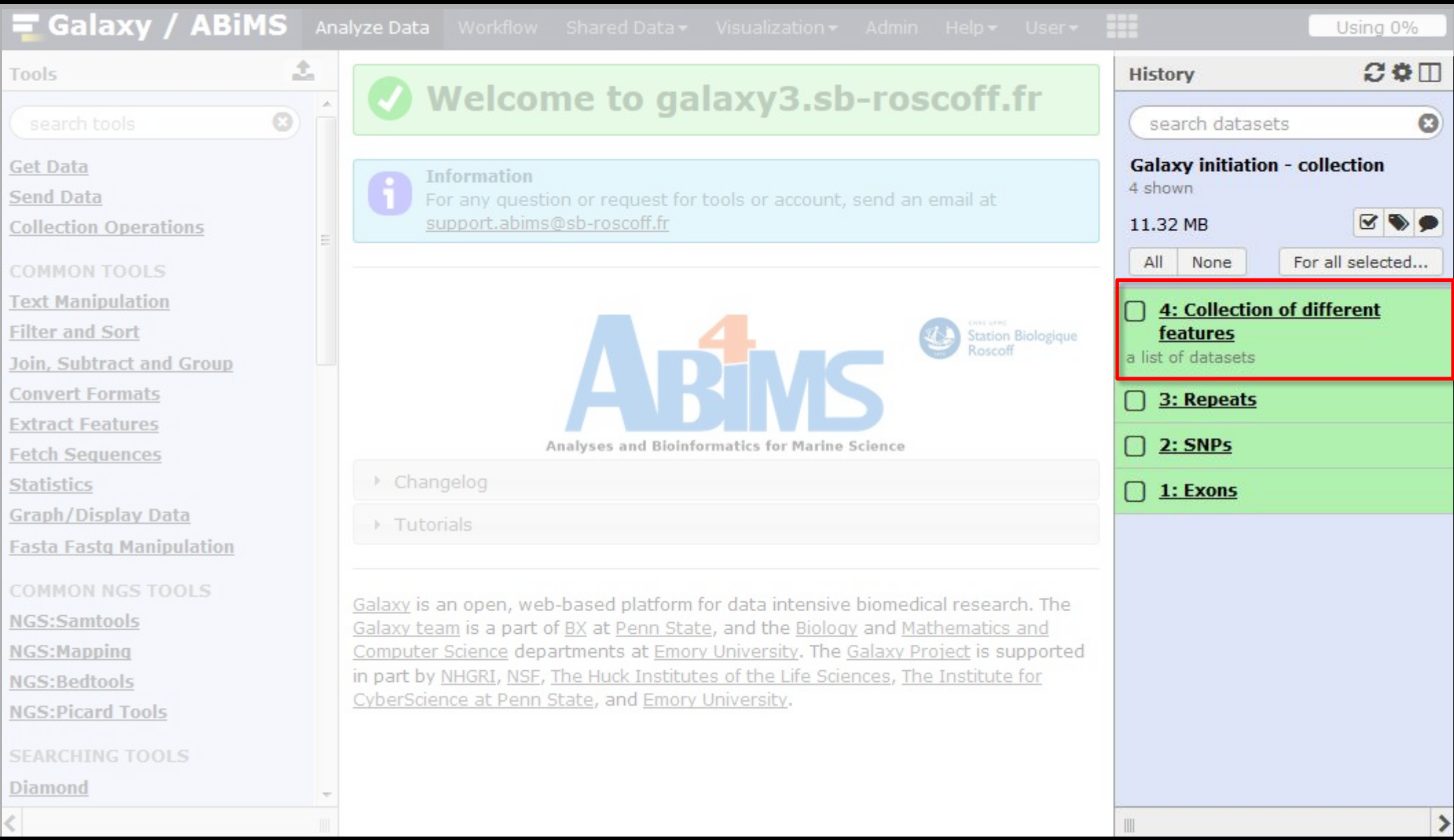
Dataset collection

Create a dataset collection



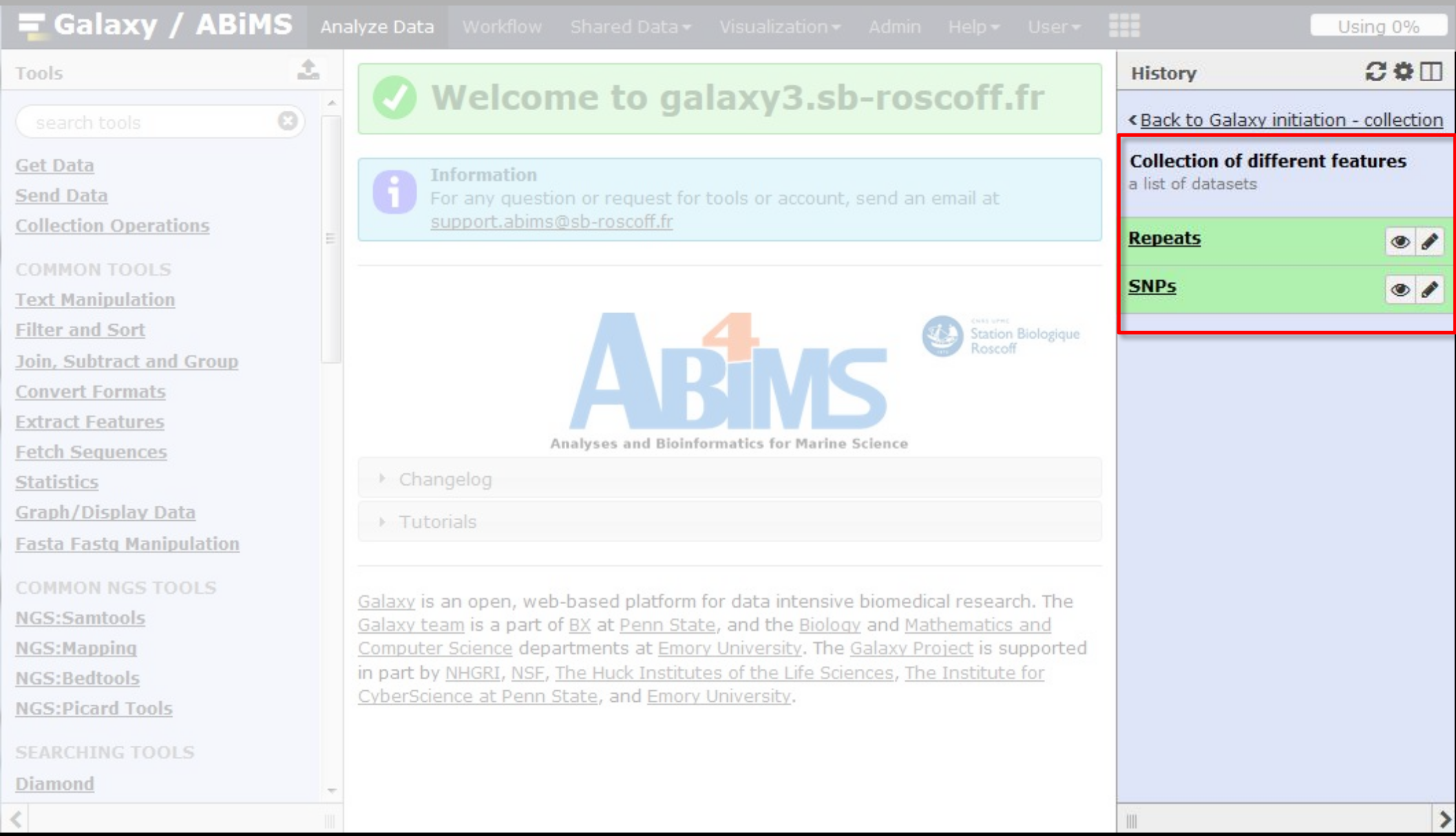
Dataset collection

Create a dataset collection



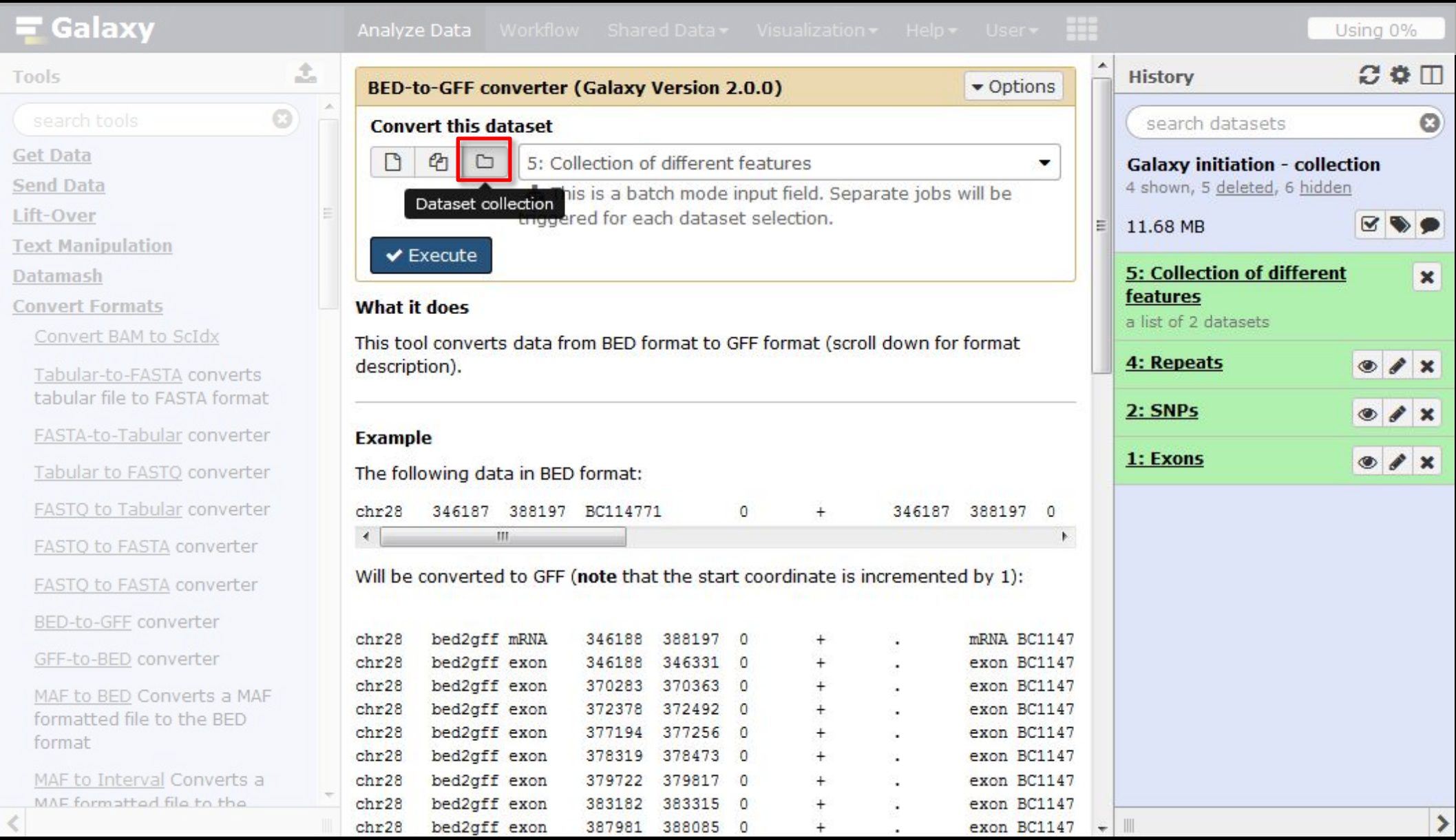
Dataset collection

Create a dataset collection



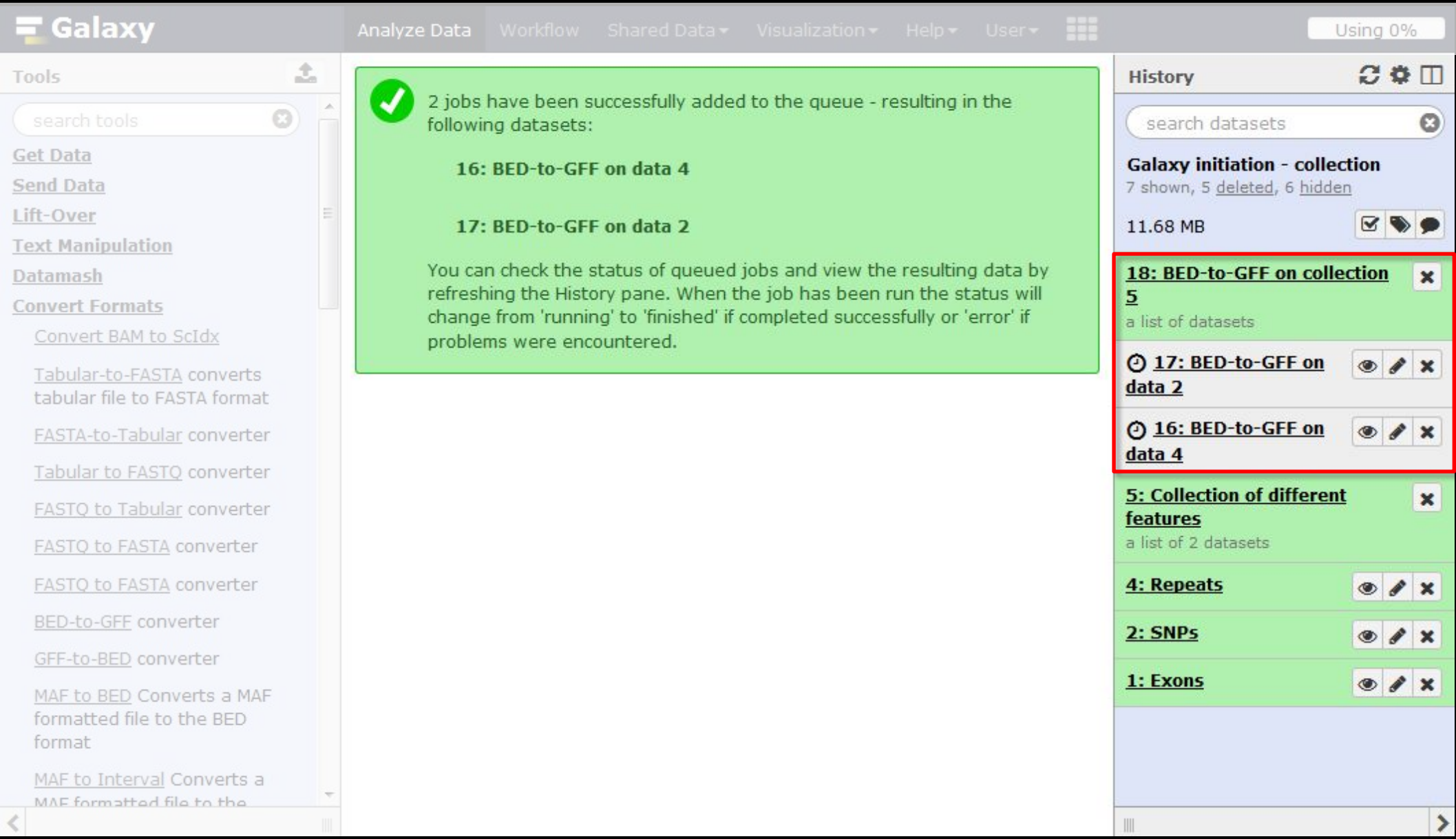
Dataset collection

Use a collection as input



Dataset collection

Use a collection as input



Hands-on

COLLECTION

Collection – Hands-on

Part 2:

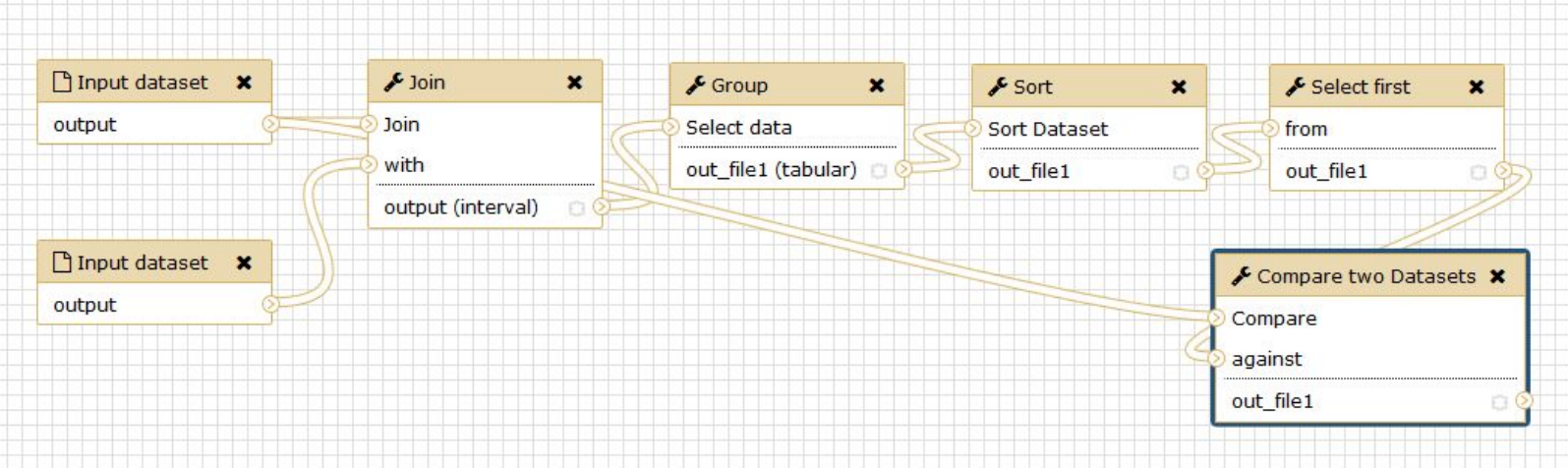
You have new sequences that you want to compare with the gene and protein databases from your favorite species (BLAST).

Case of Drosophila melanogaster (dm)

1. Create new history
2. Import CDS and peptide sequences databases
3. Import query sequences
4. Make BLAST databases
5. Run BLAST against the CDS database
6. Run BLAST against the protein database
7. Create a dataset list with all the query sequences
8. Run BLAST against the CDS database on the dataset list

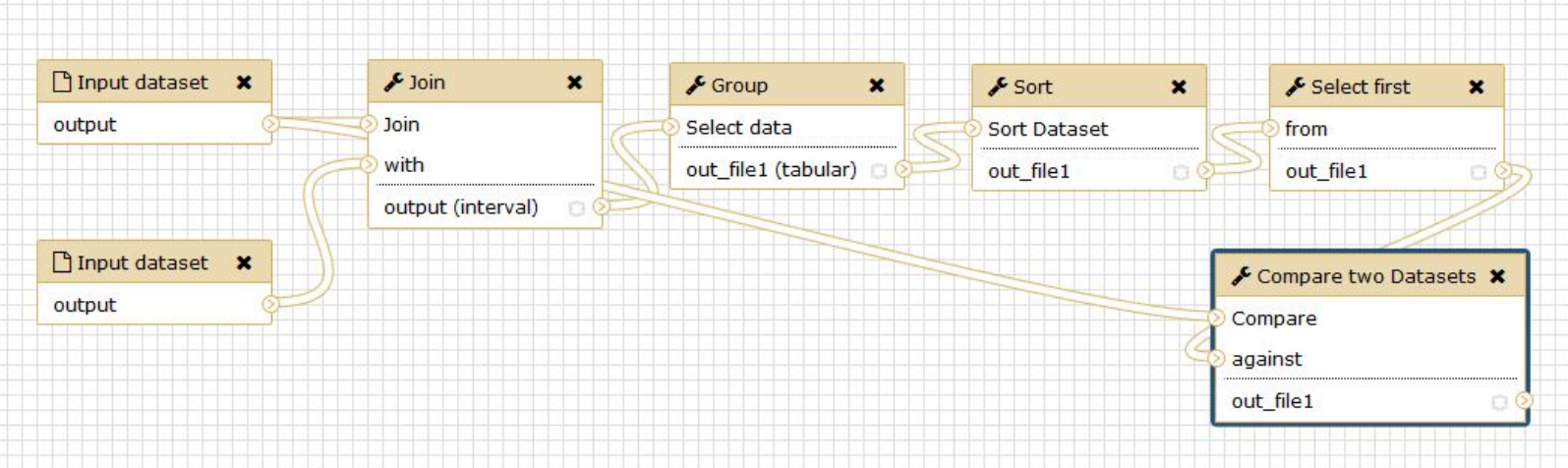
WORKFLOW

Workflow



- A workflow is a sequence of tool operations and parameters
- A workflow is built to be replayed (more or less strict)

Workflow

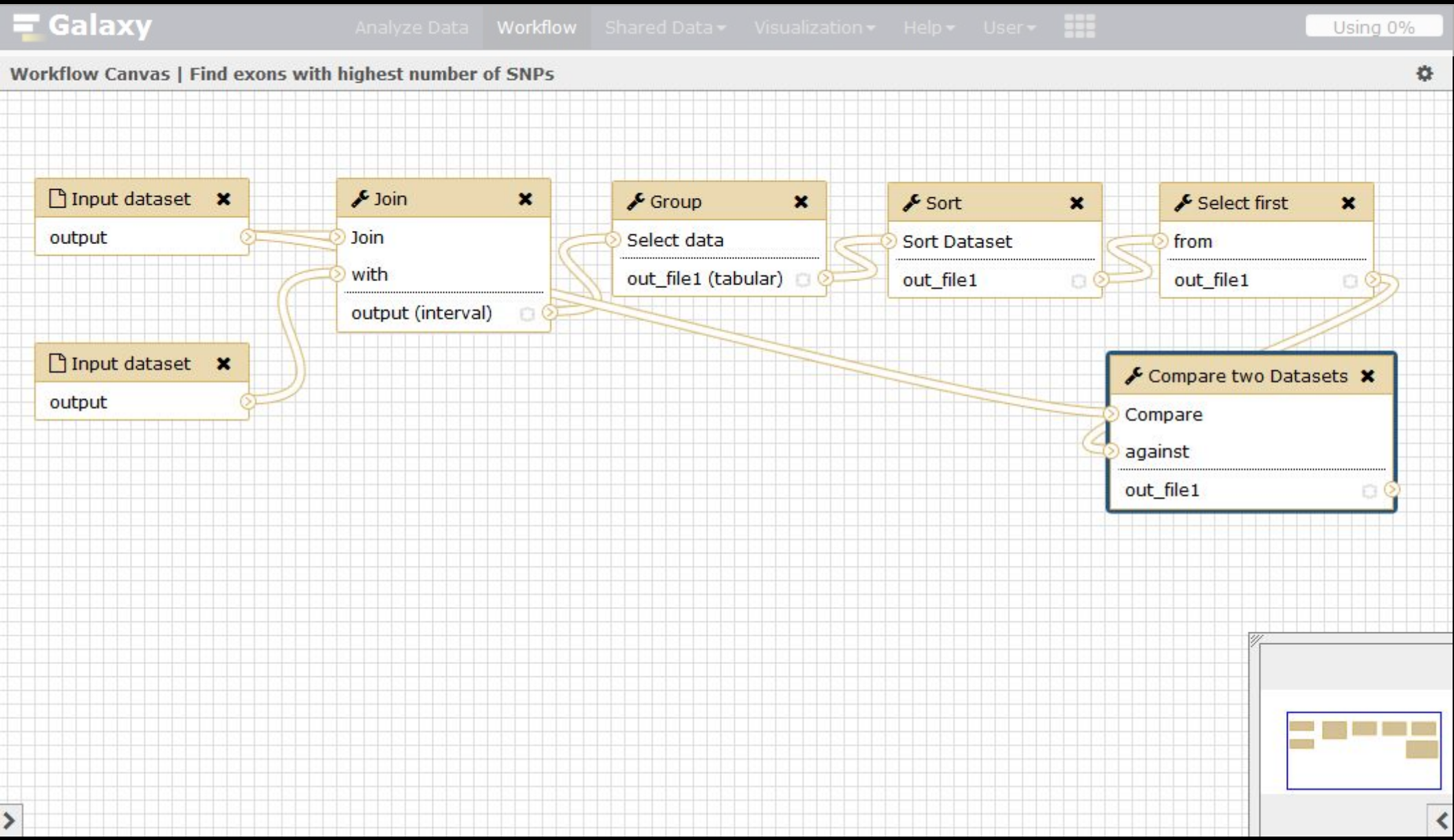


Why would you want to create workflows?

- 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)
- Share workflows for publication and with the community

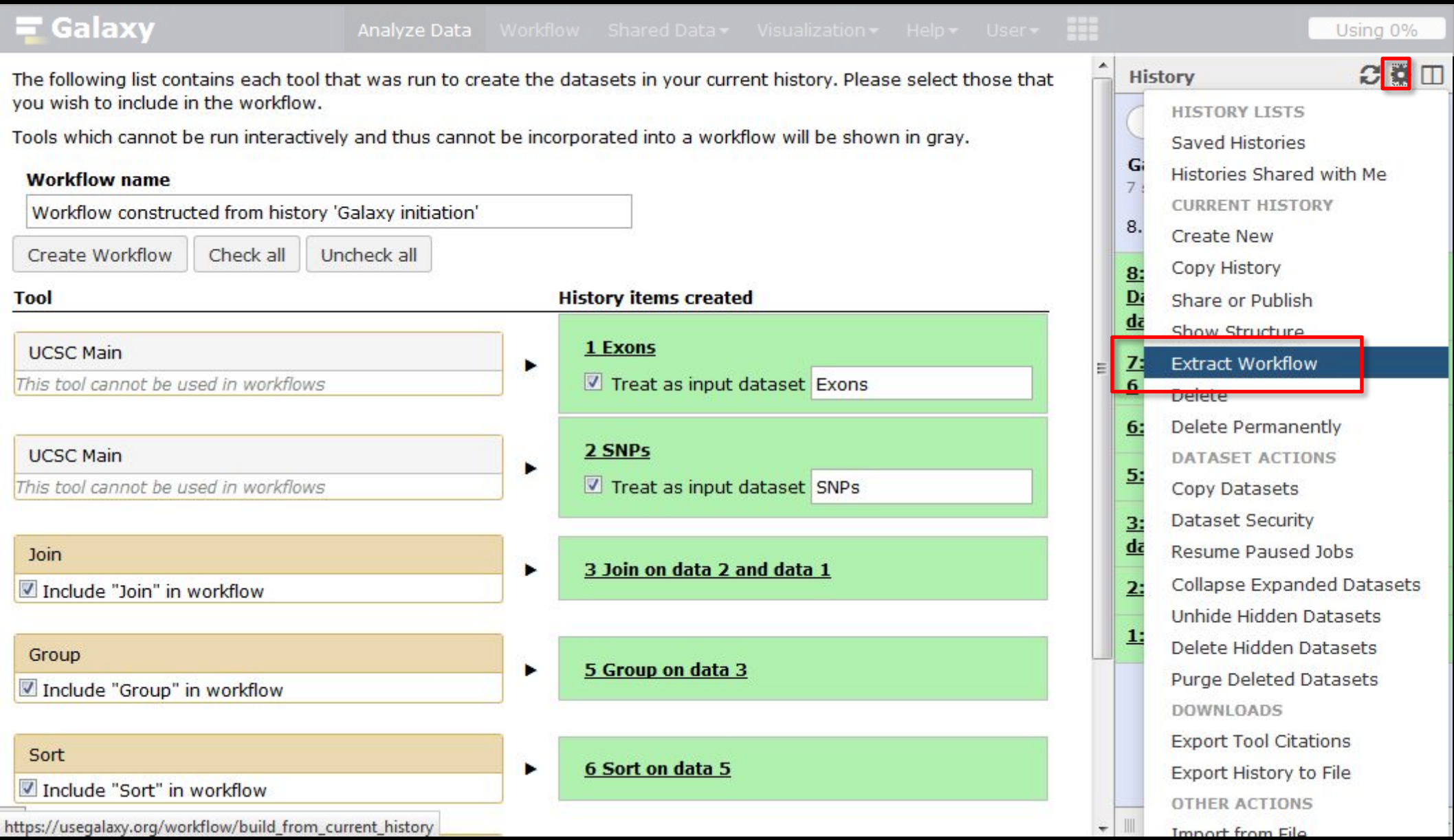
Workflow

Our workflow with Galaxy



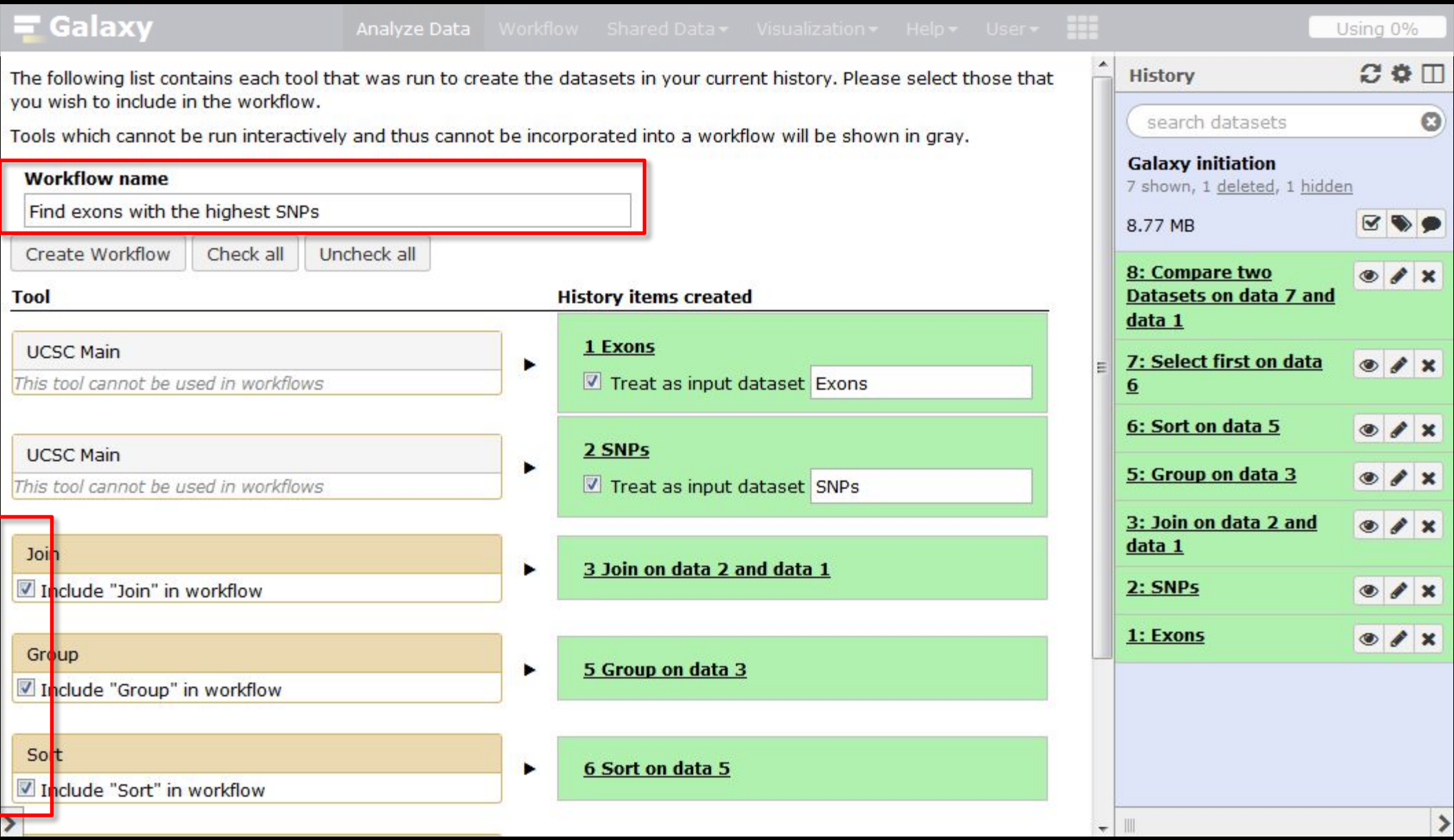
Workflow

From history



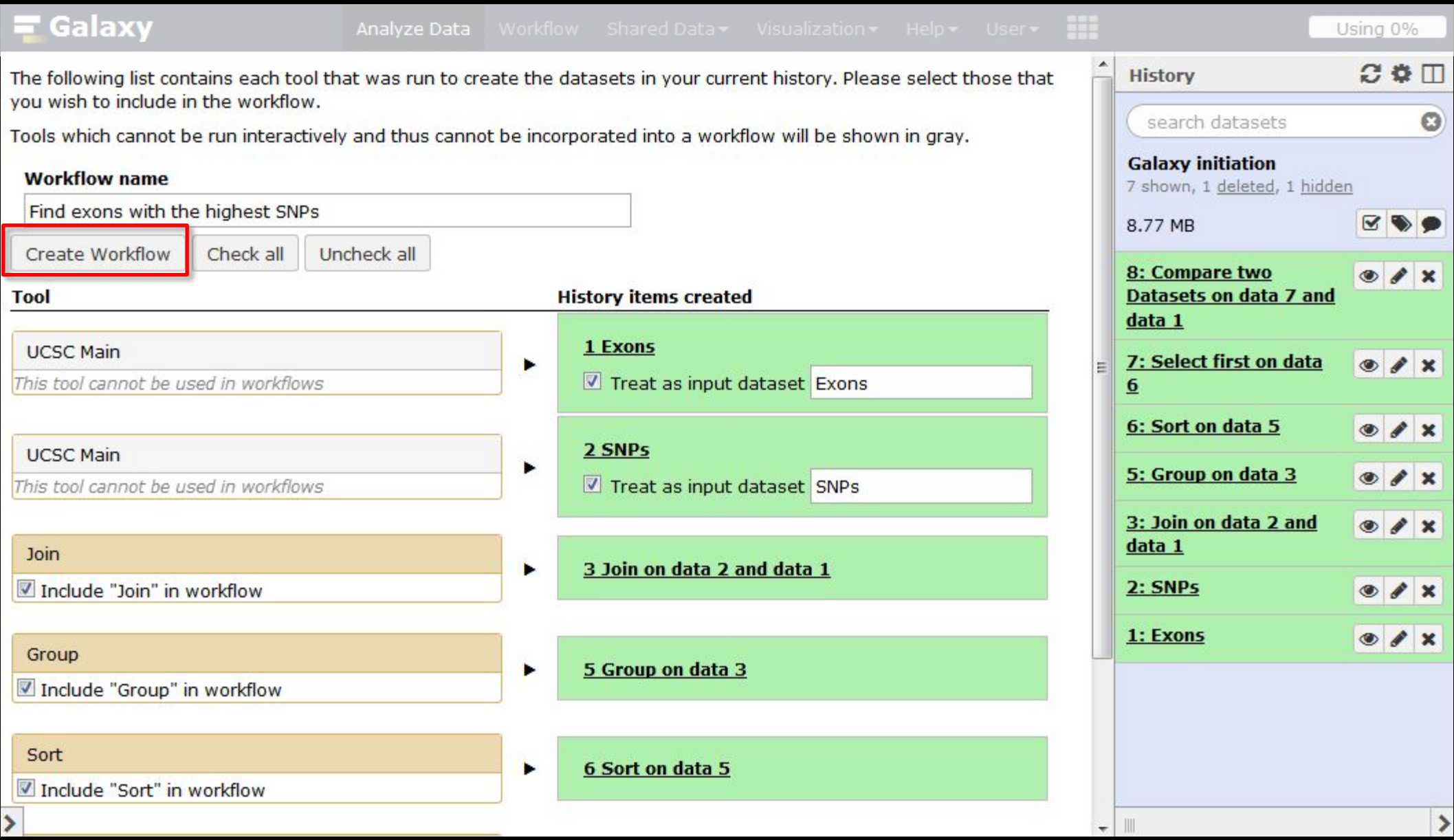
Workflow

From history



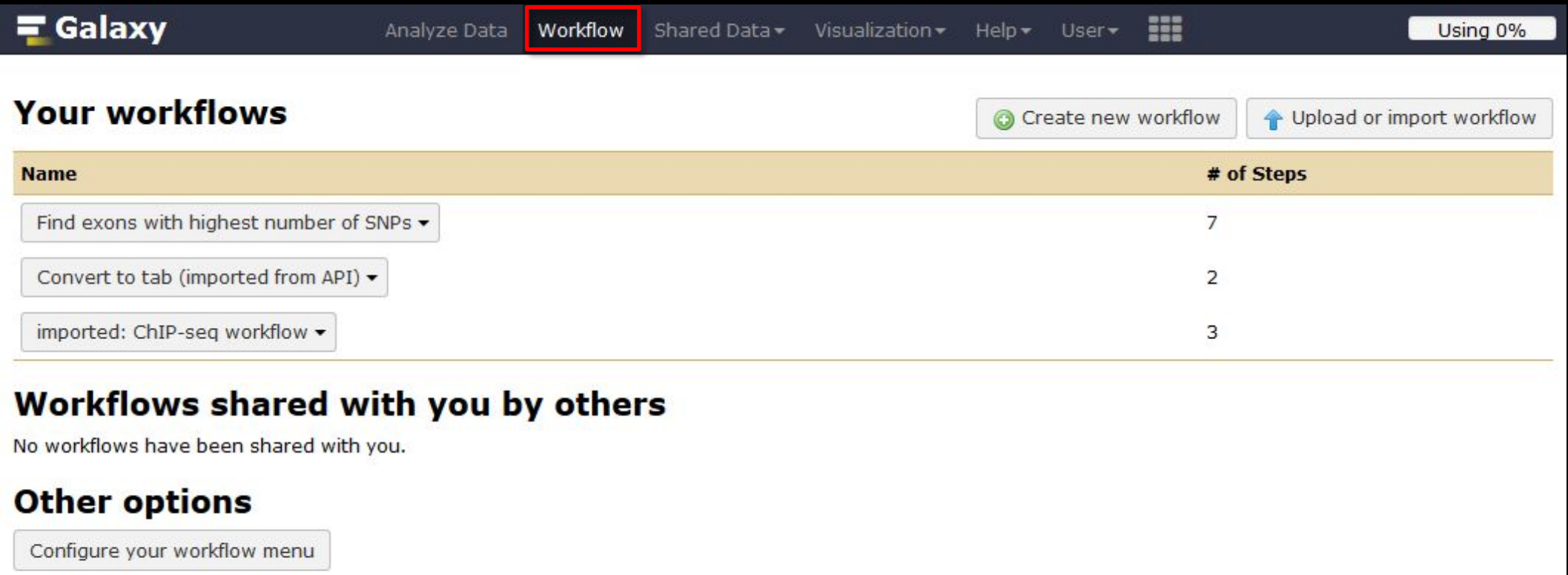
Workflow

From history



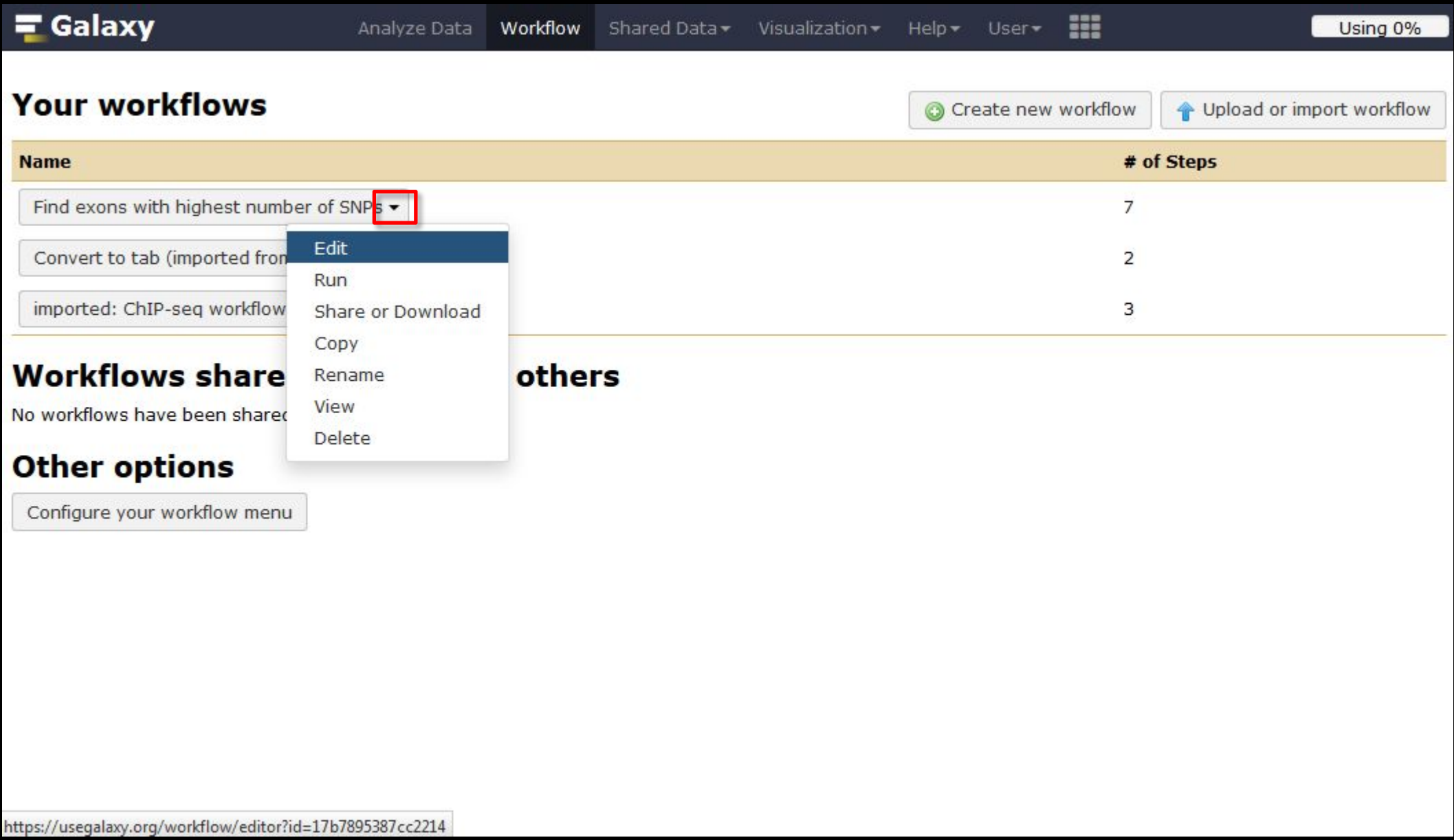
Workflow

Workflow manager



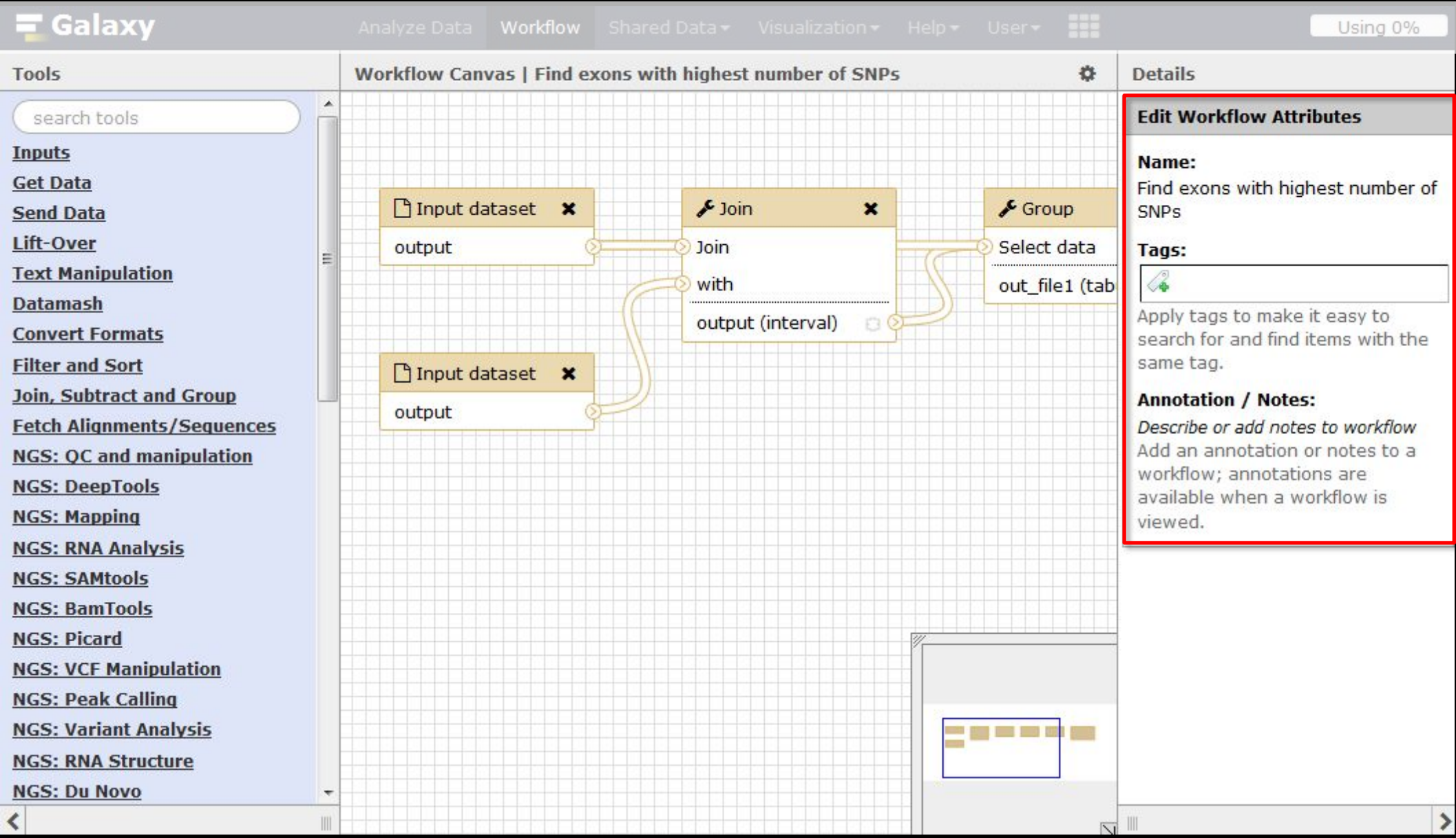
Workflow

Workflow manager



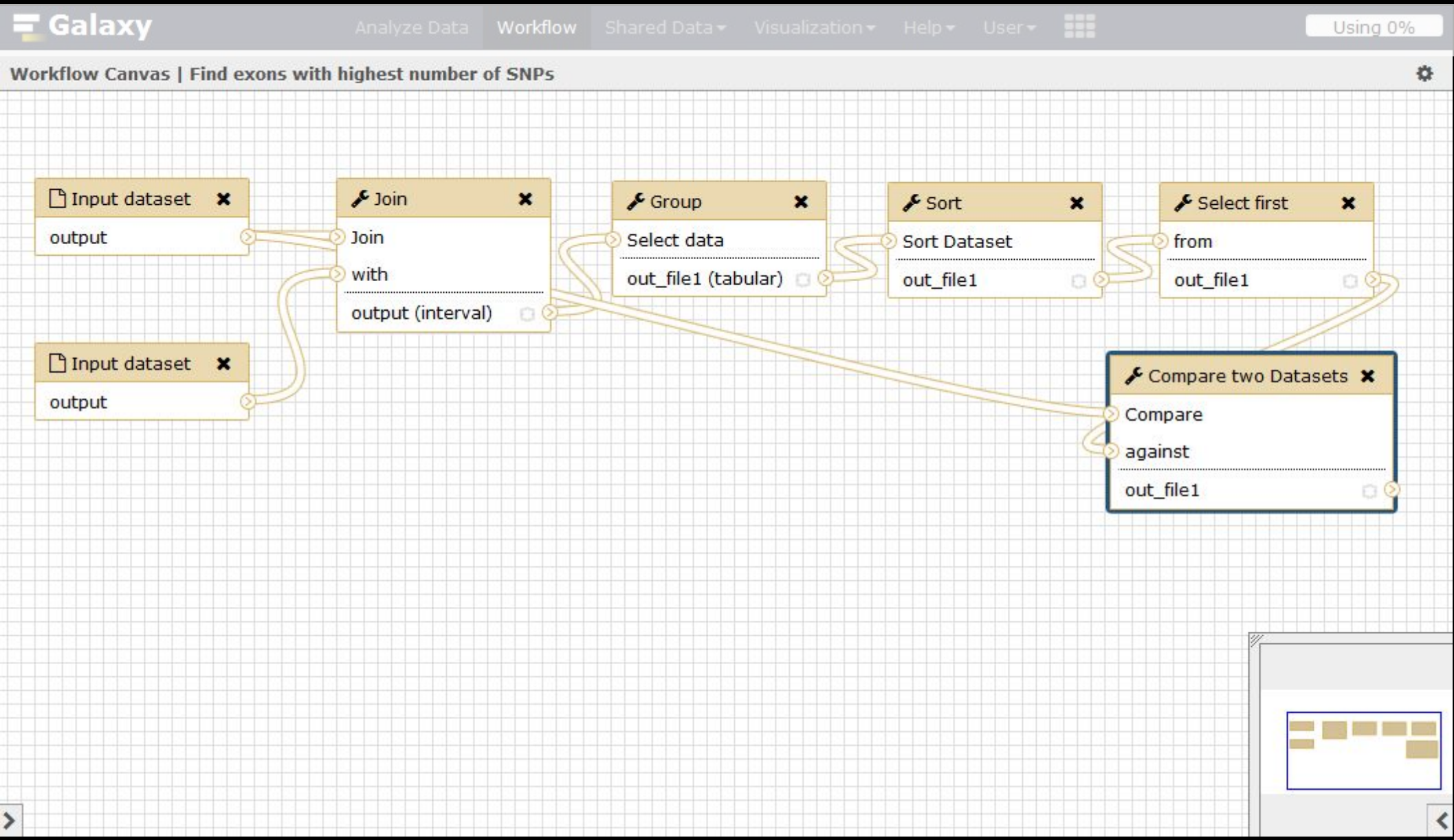
Workflow

Edit a workflow: add tags and annotation



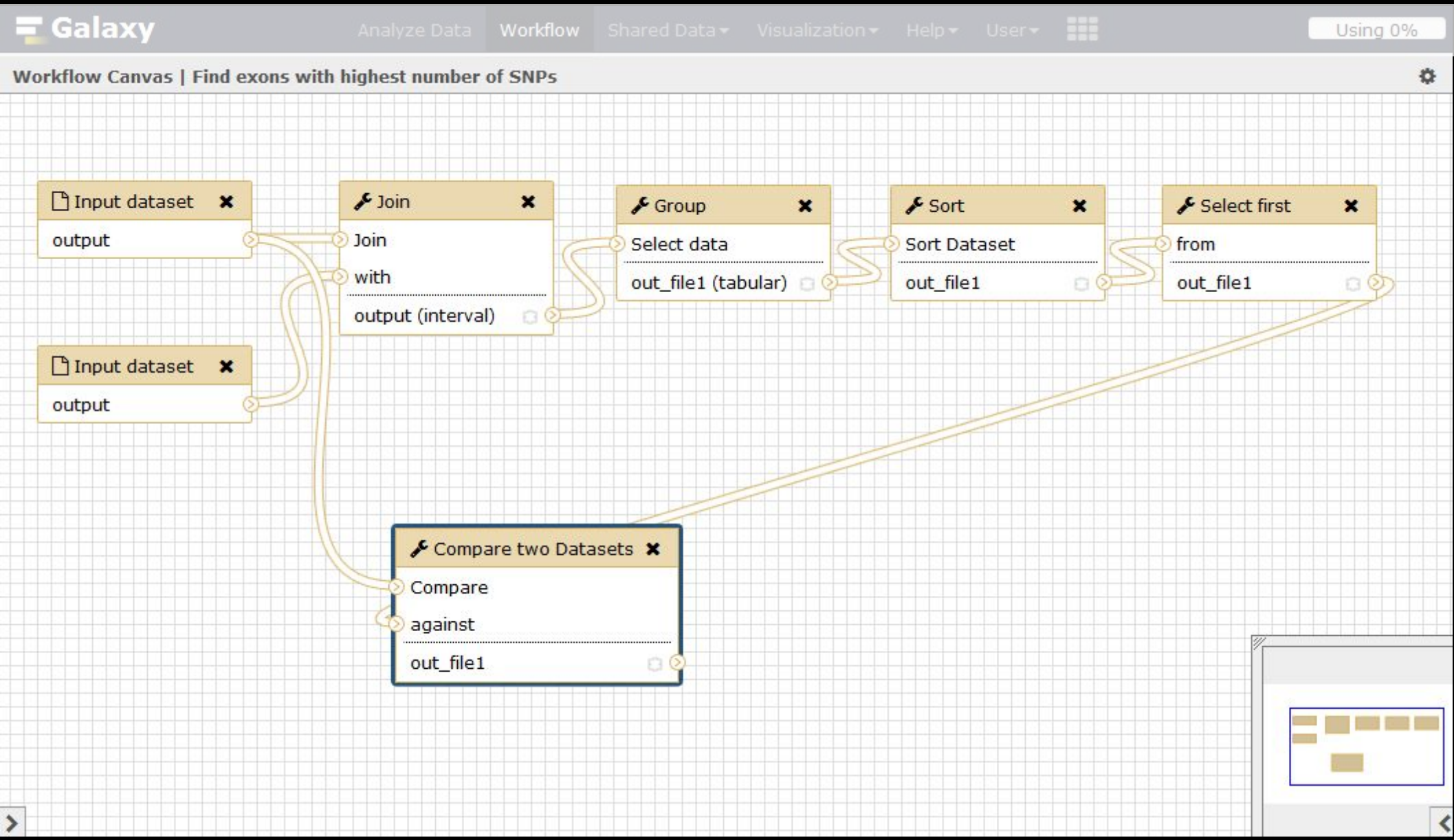
Workflow

Edit a workflow



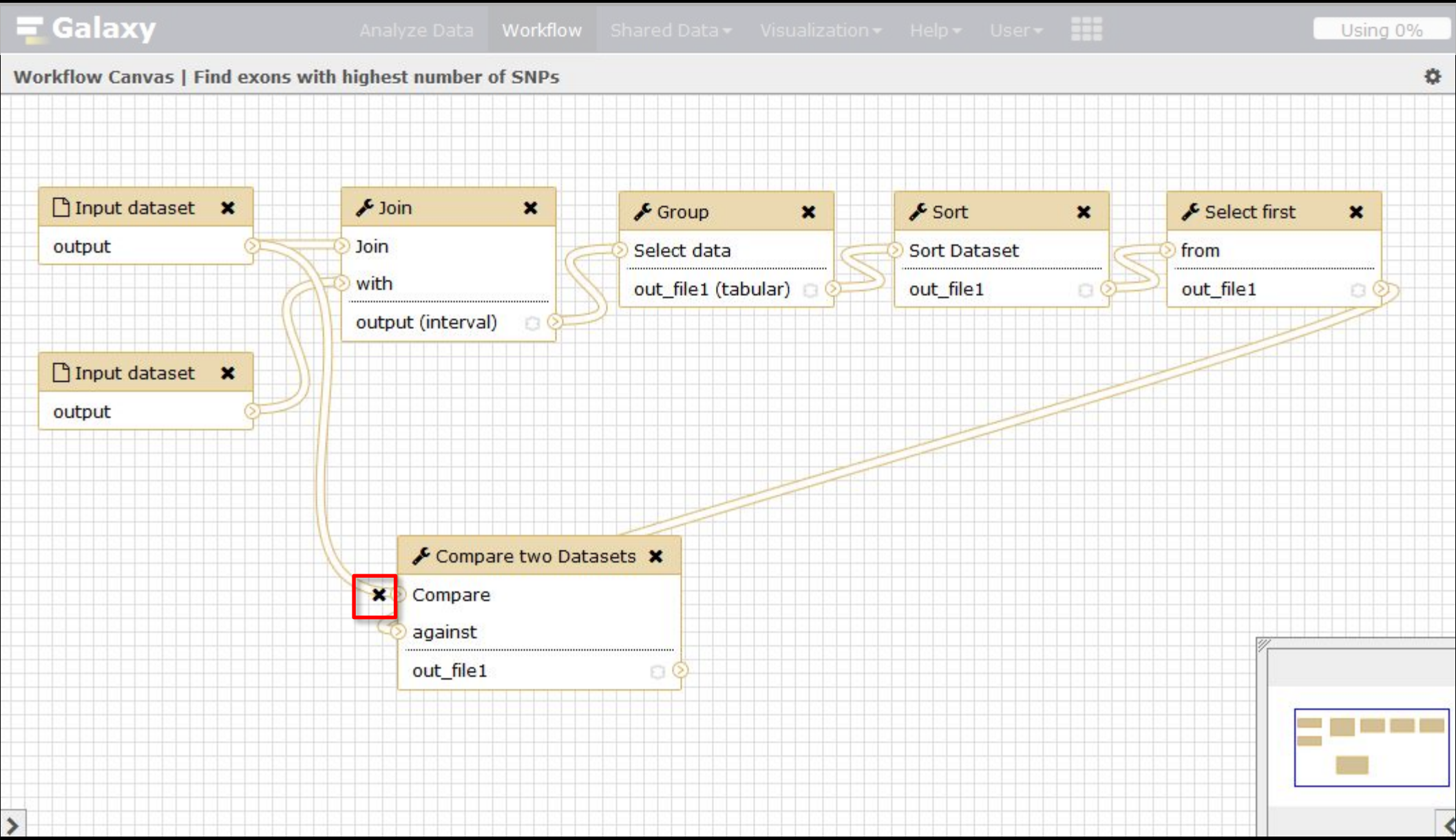
Workflow

Edit a workflow: drag and drop



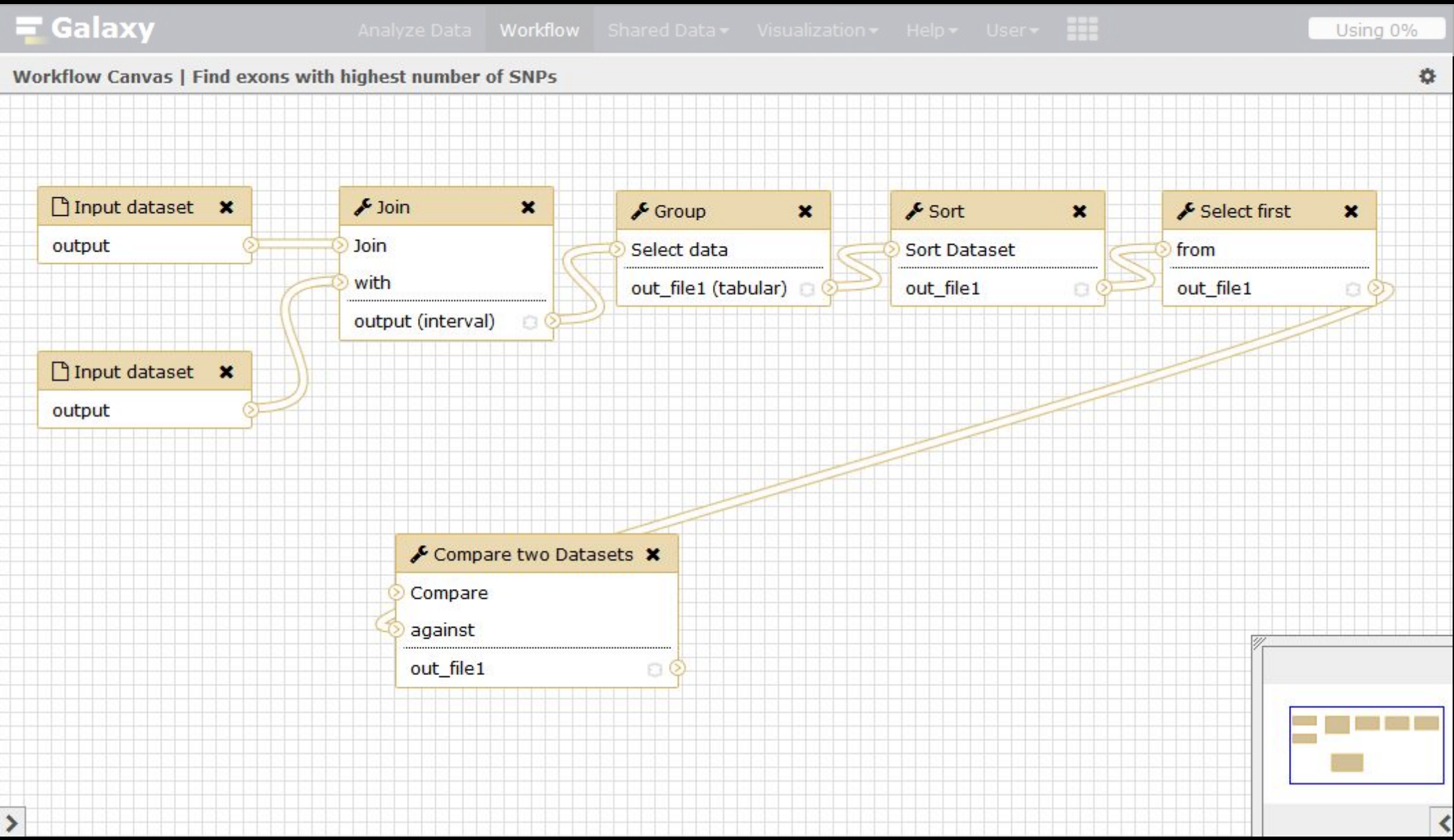
Workflow

Edit a workflow: delete a noodle



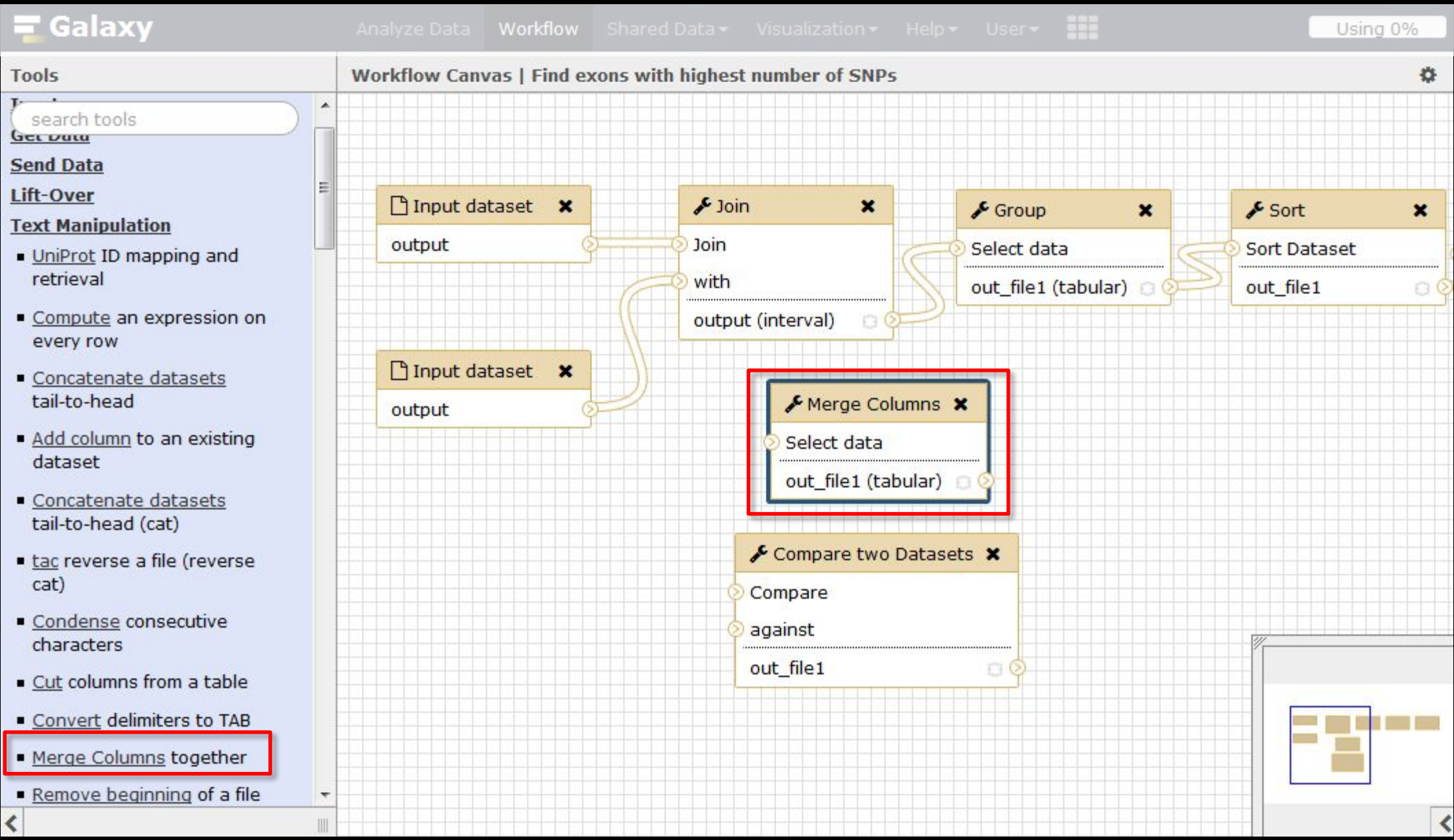
Workflow

Edit a workflow: delete a noodle



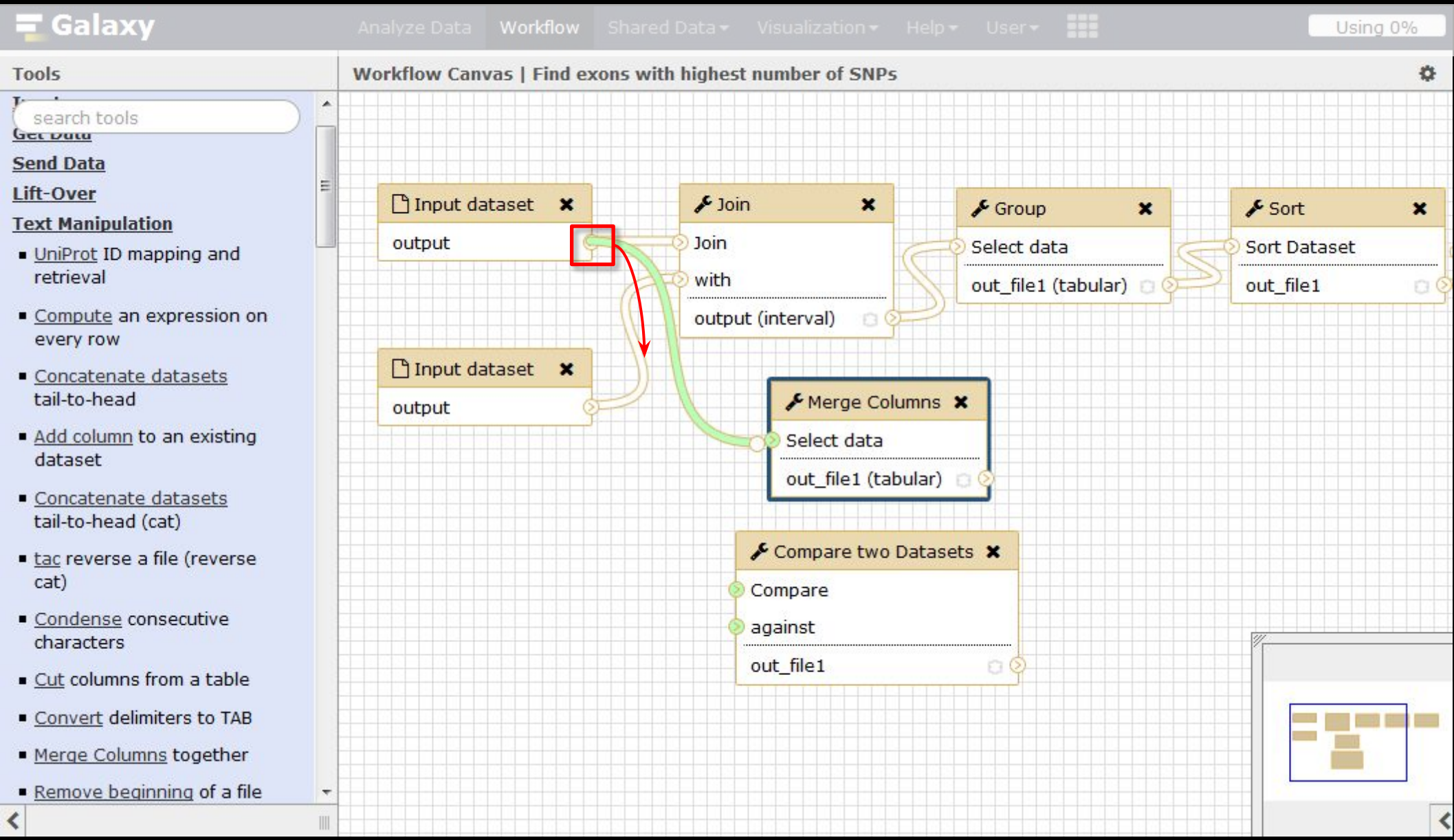
Workflow

Edit a workflow: add a tool



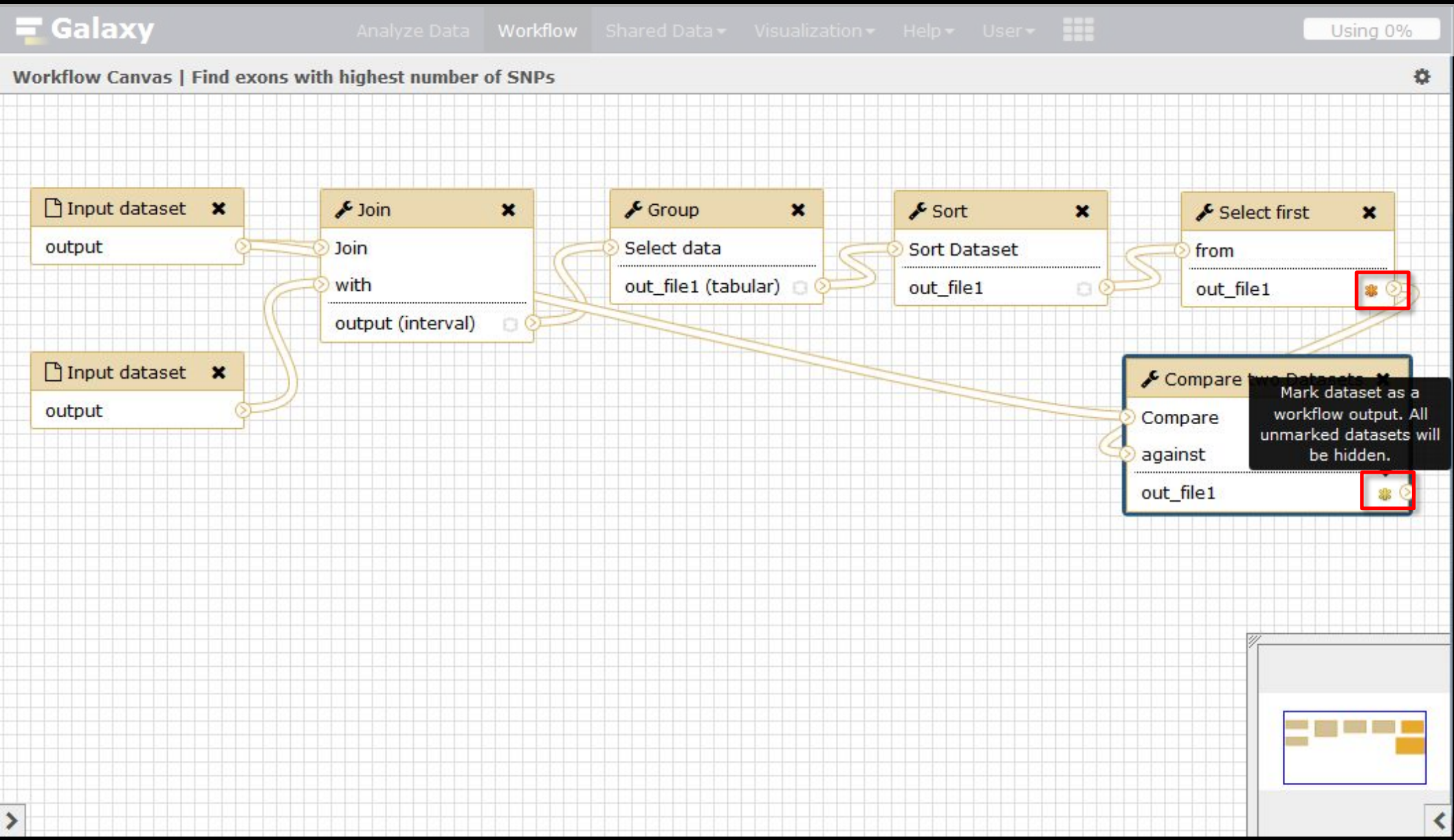
Workflow

Edit a workflow: add a noodle



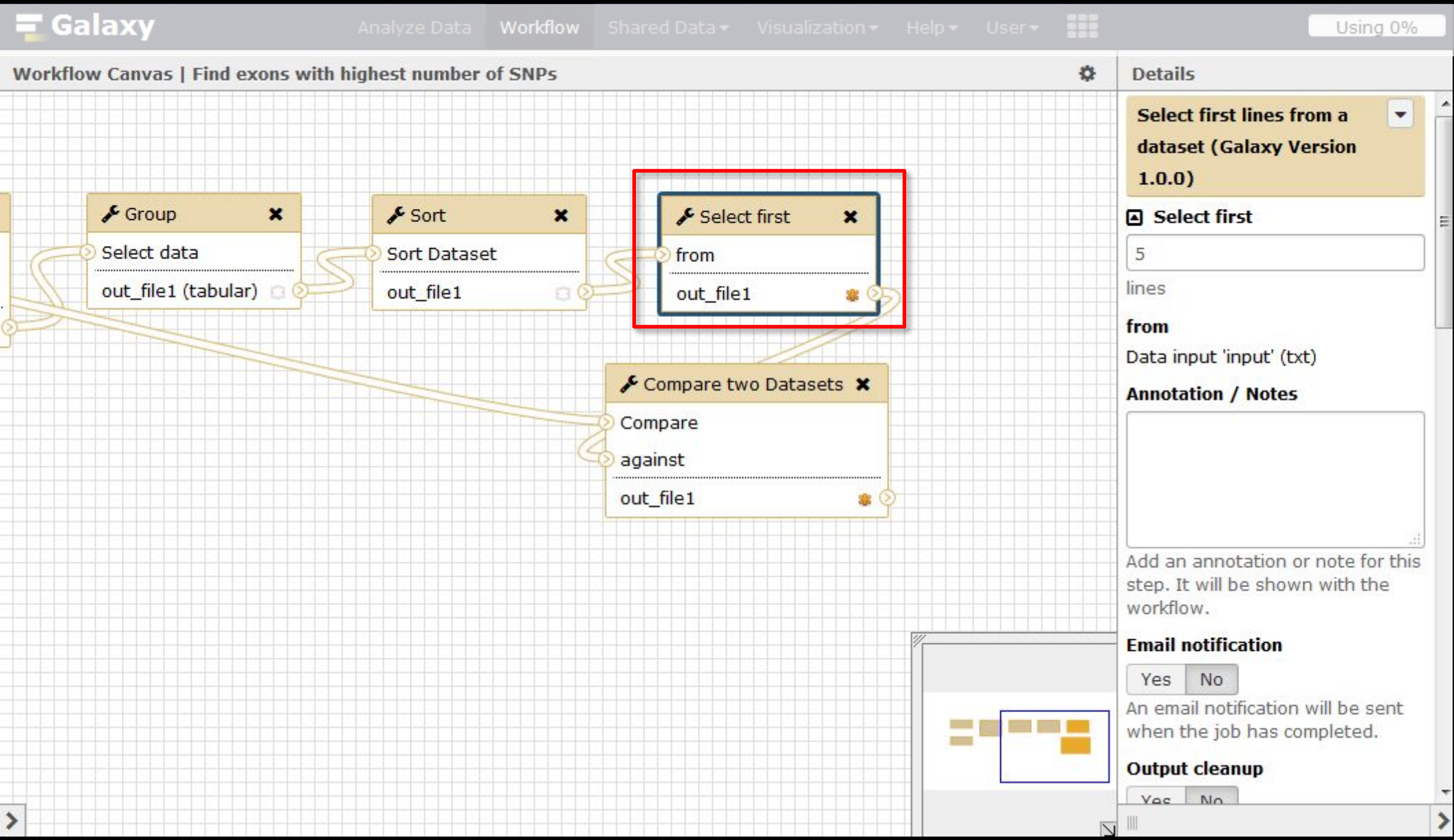
Workflow

Edit a workflow: hide intermediate steps



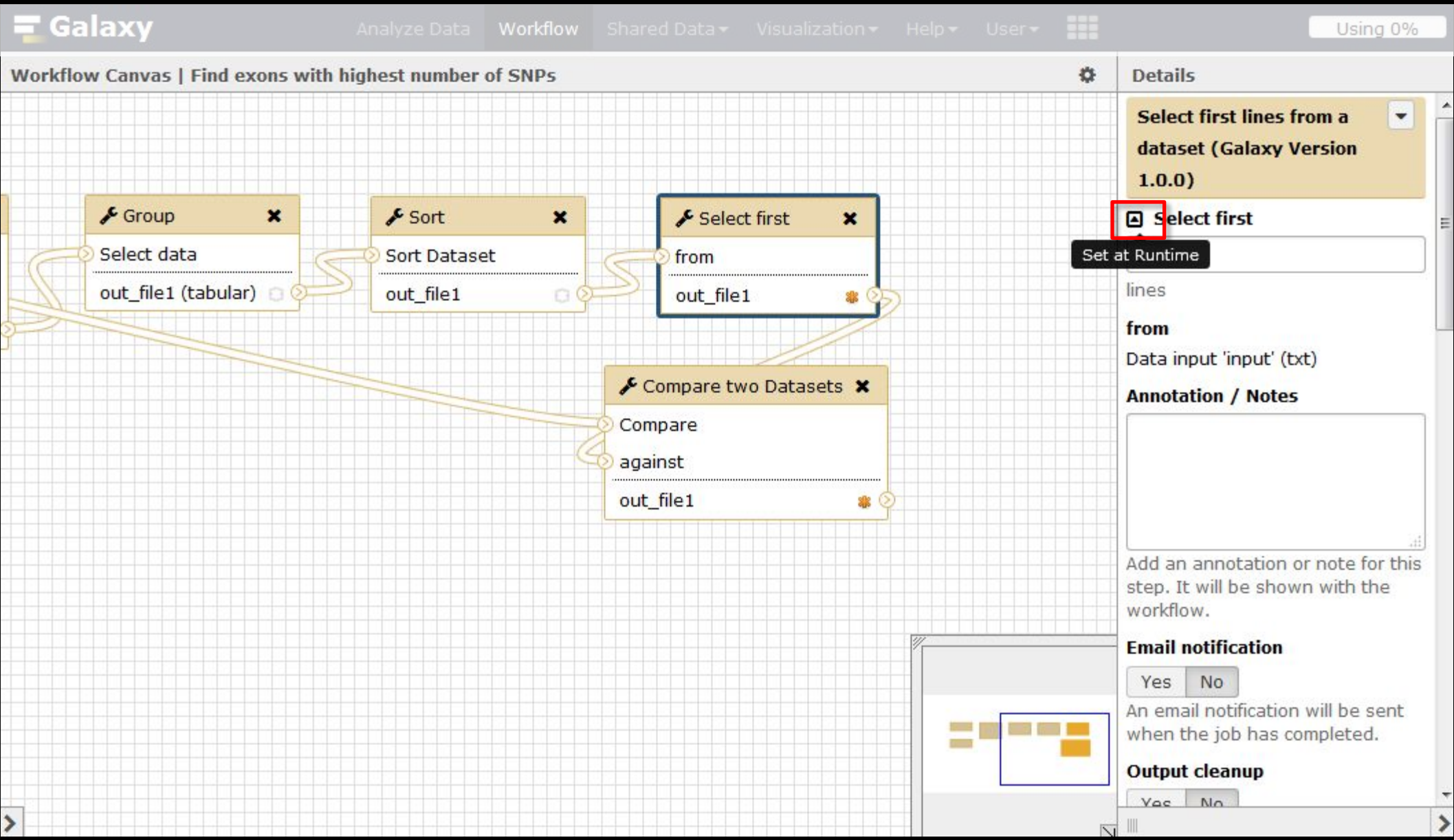
Workflow

Edit a workflow: set or release a parameter



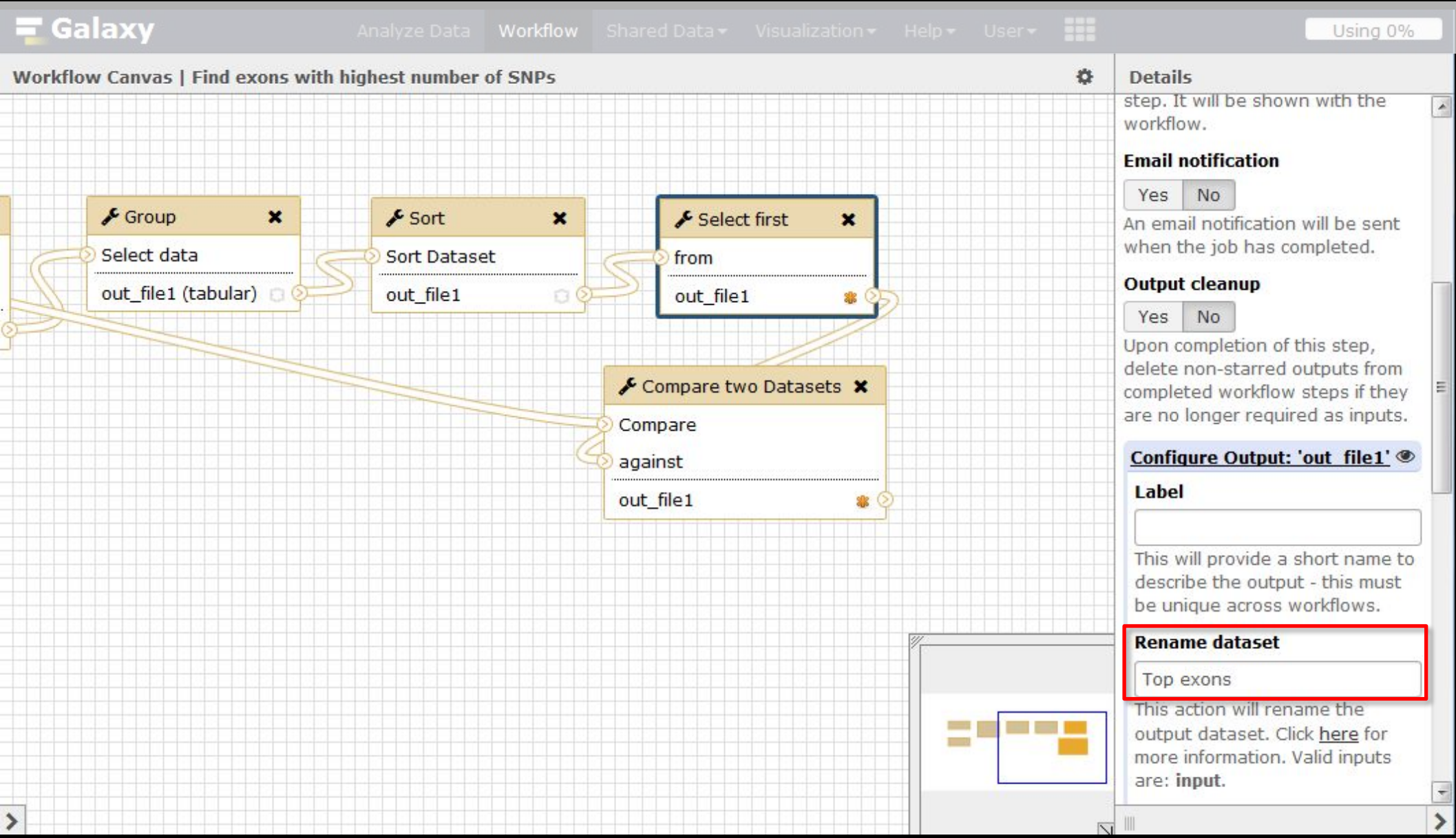
Workflow

Edit a workflow: set or release a parameter



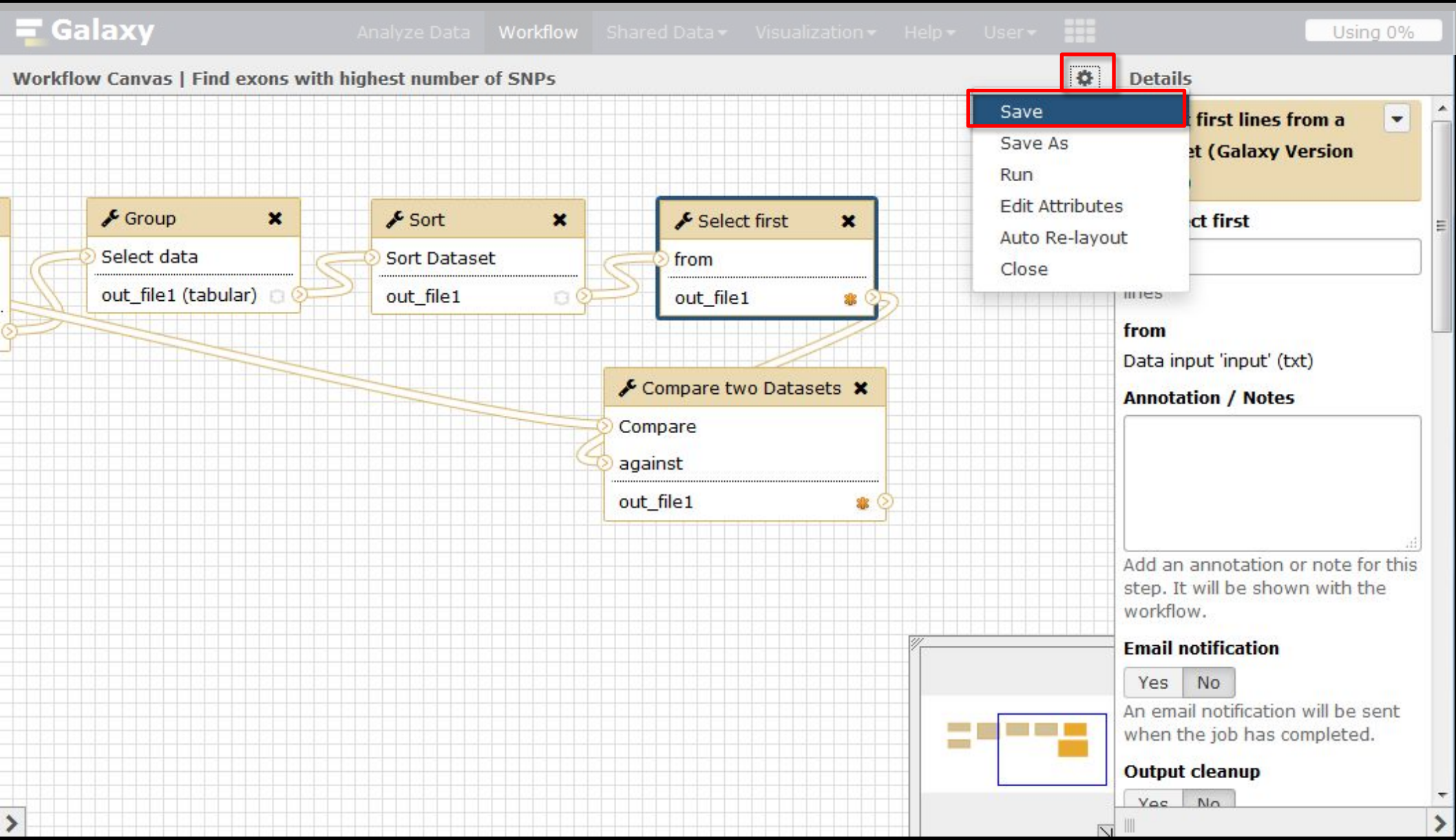
Workflow

Edit a workflow: rename the outputs



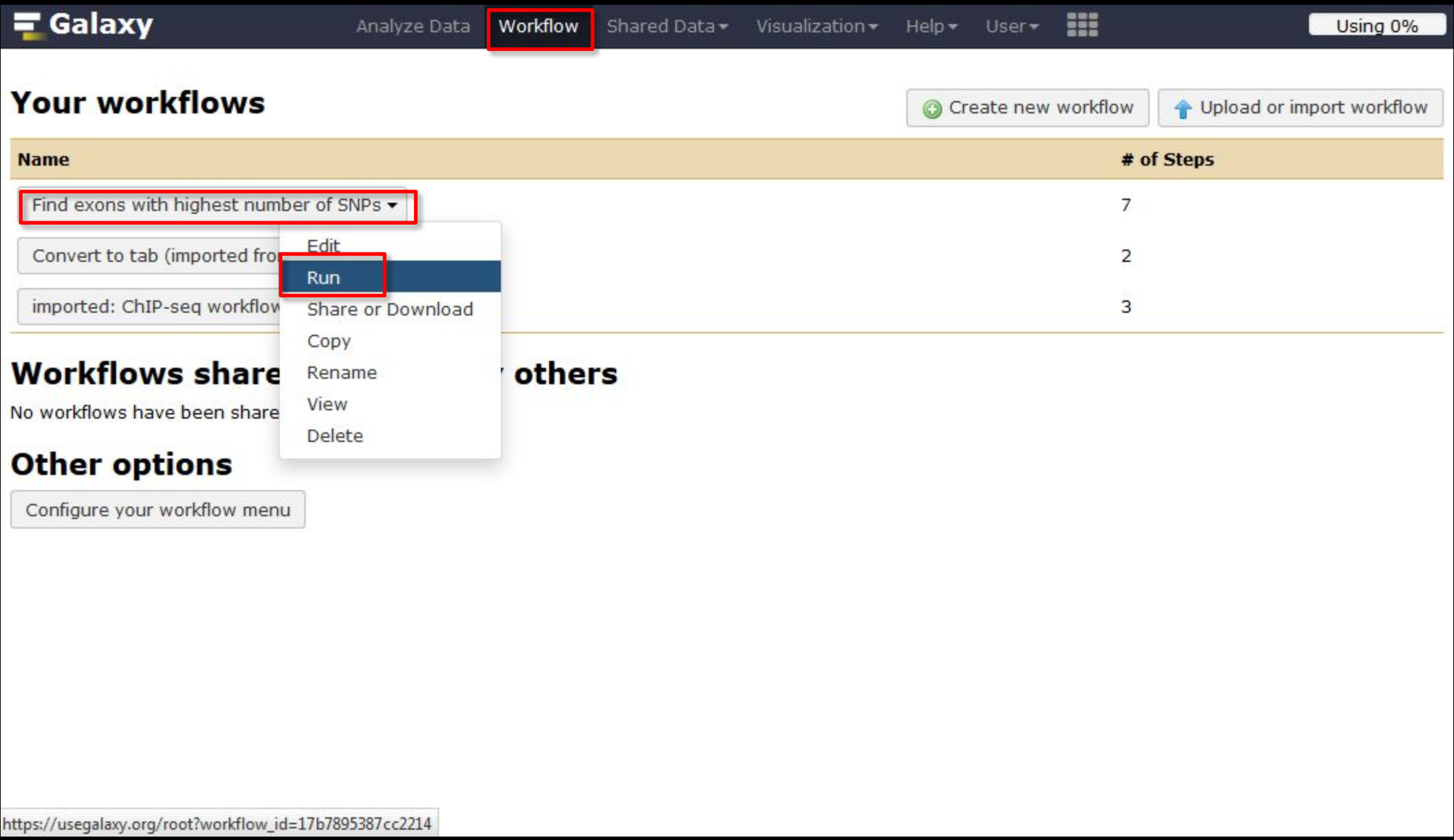
Workflow

Save



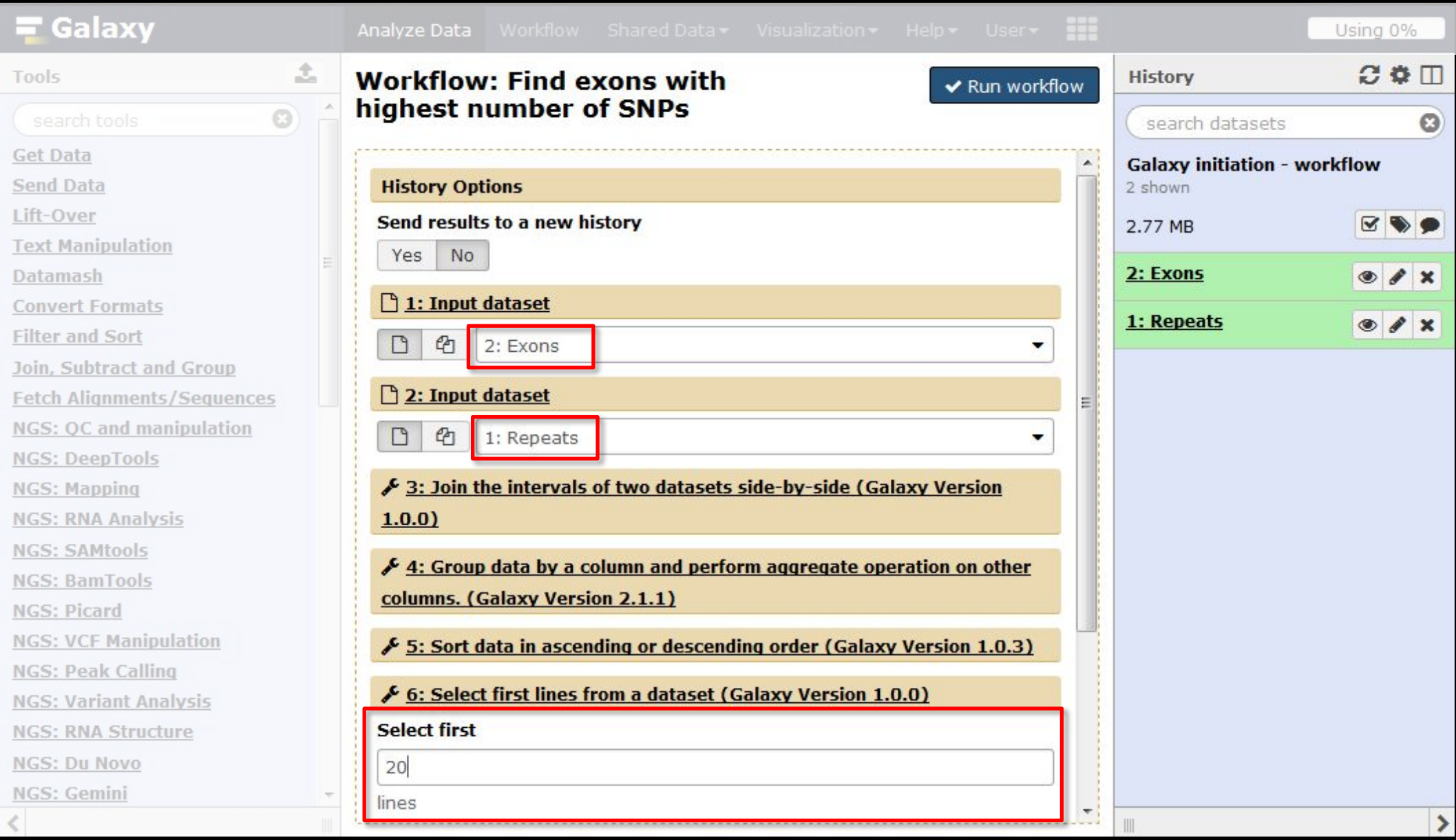
Workflow

Run a workflow



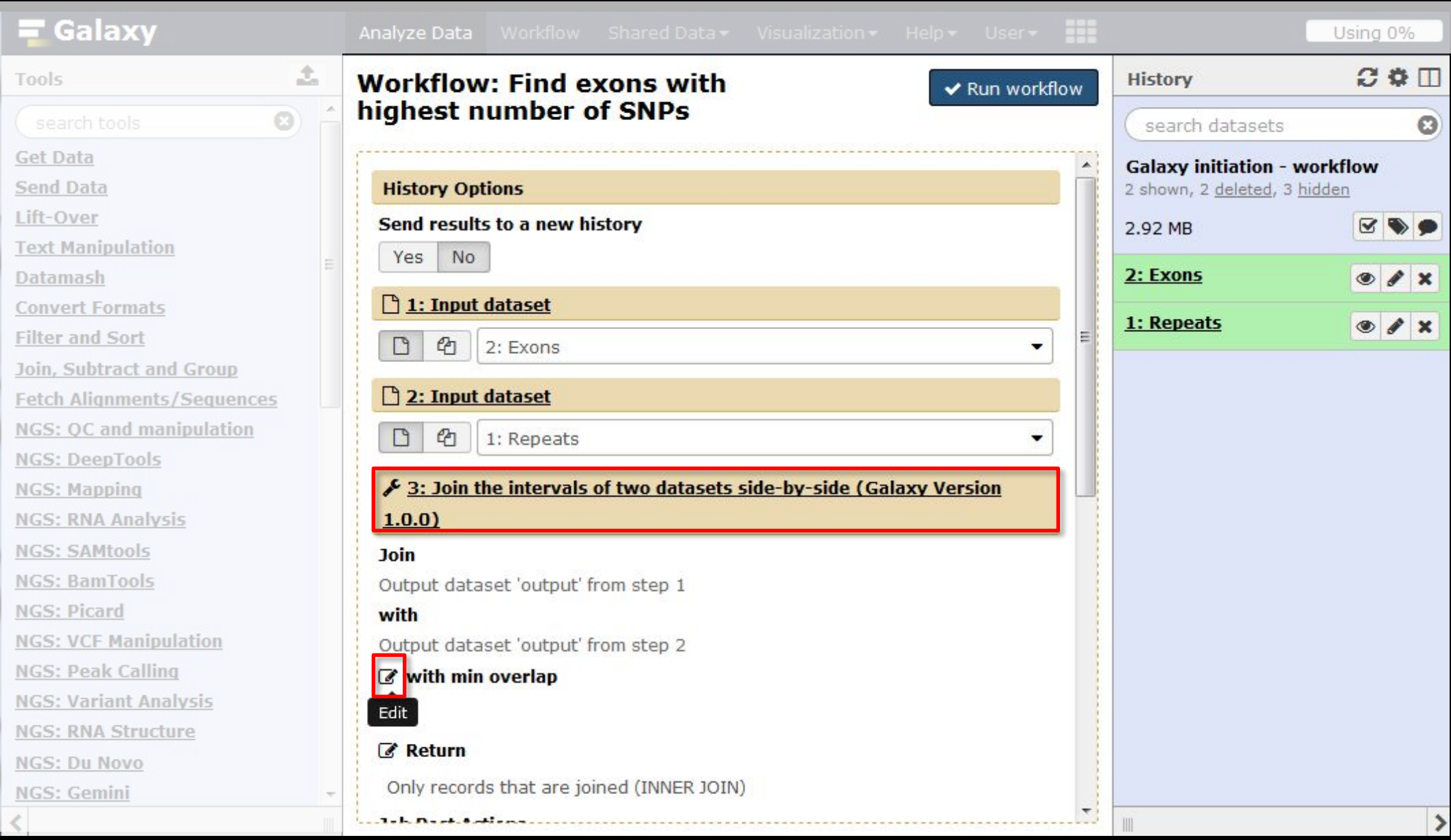
Workflow

Run a workflow



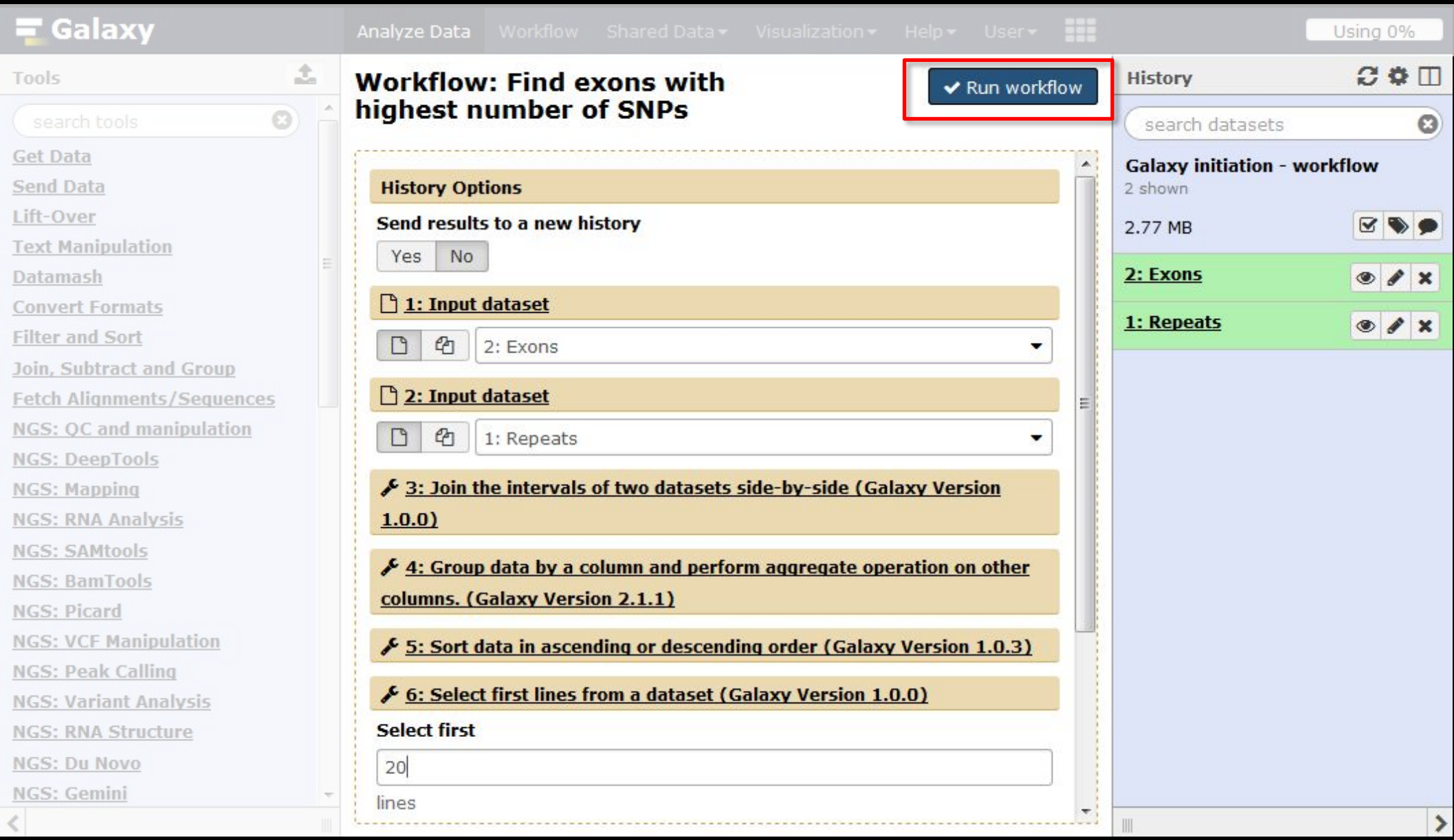
Workflow

Run a workflow



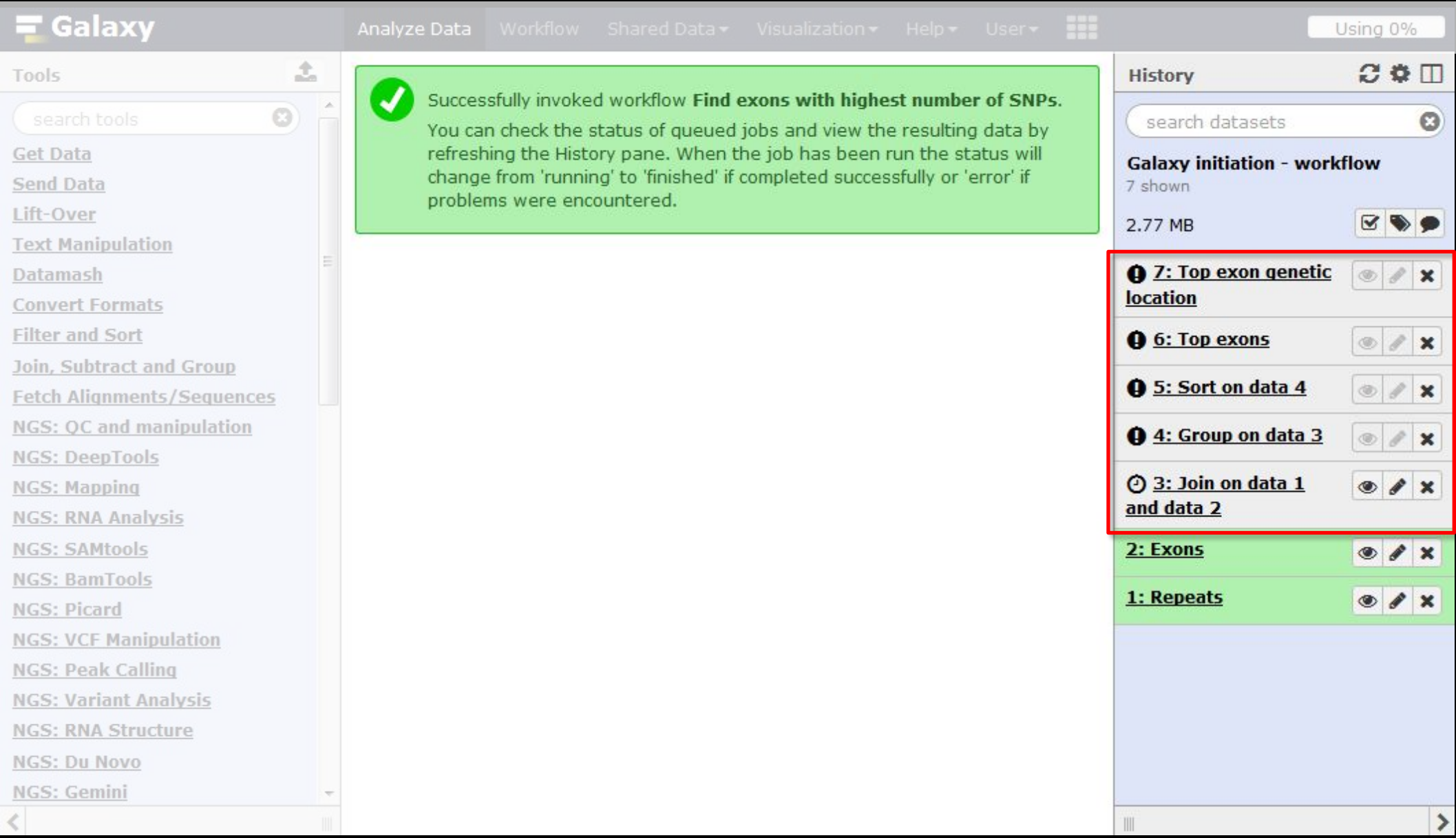
Workflow

Run a workflow



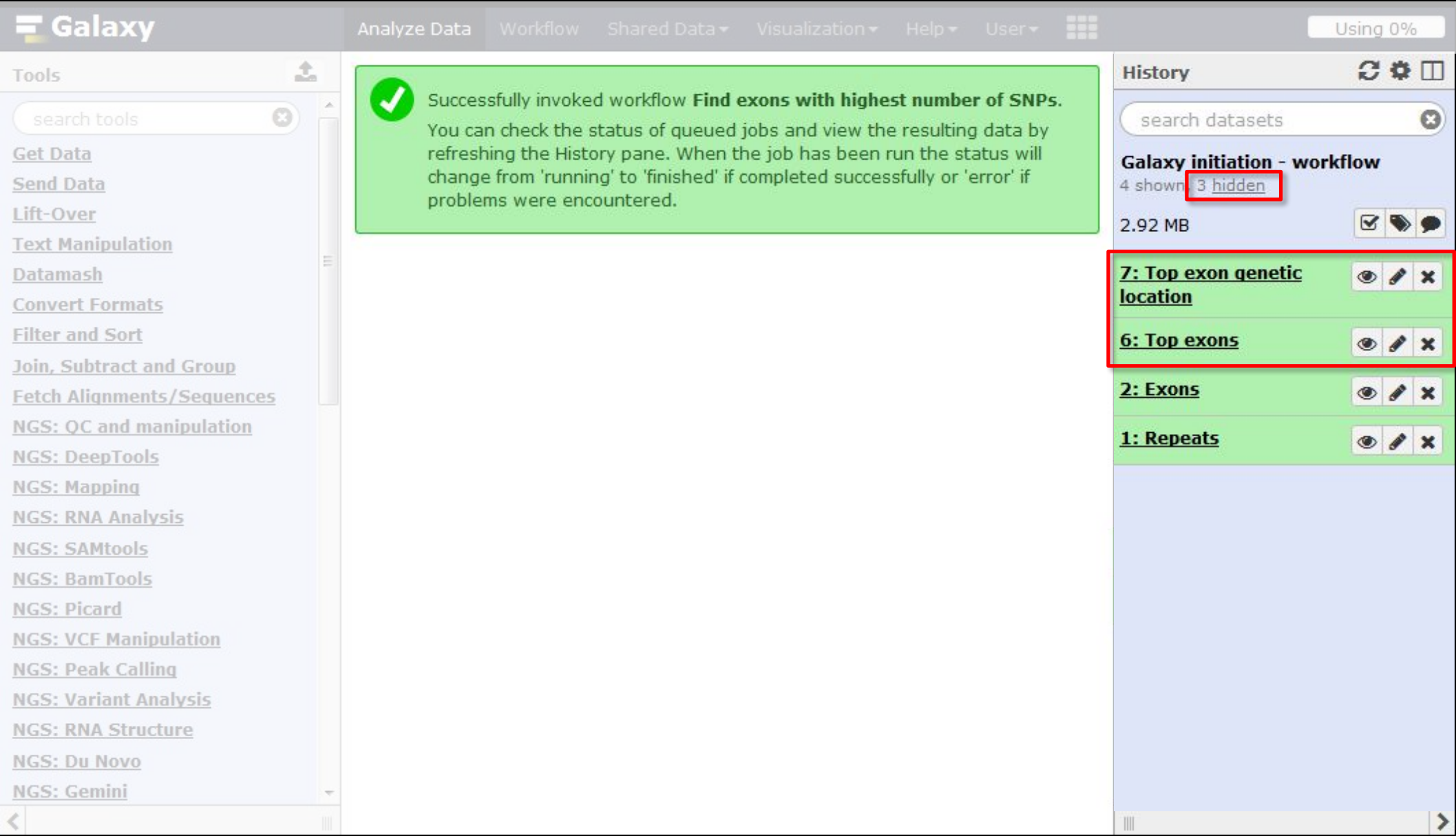
Workflow

Run a workflow



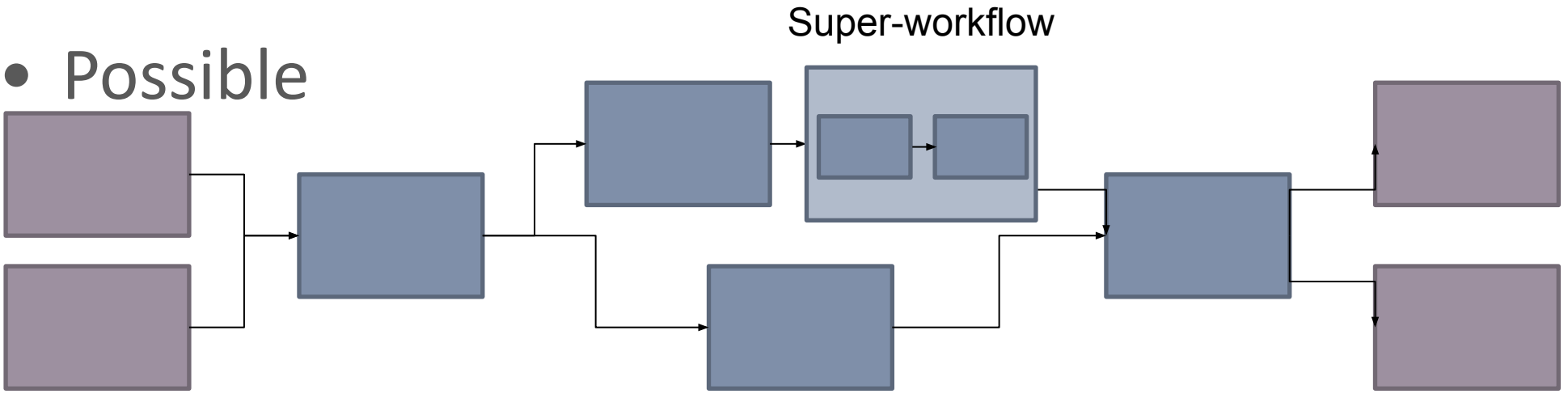
Workflow

Run a workflow

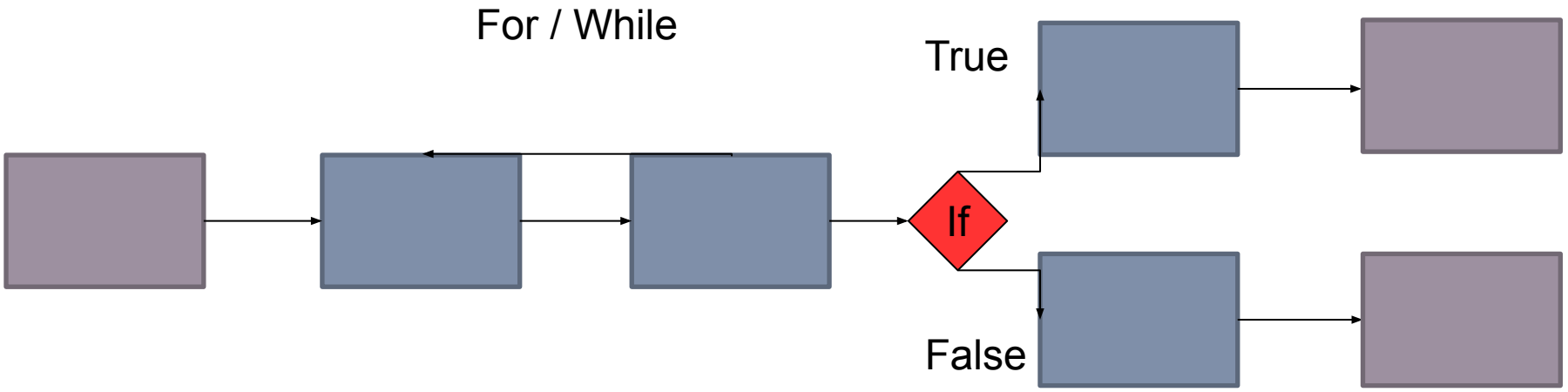


Workflow

- Possible



- Impossible (until now)



SHARE

biologist ↔ biologist

- Sharing histories or datasets
  - With or without linked workflow

bioanalyst ↔ biologist

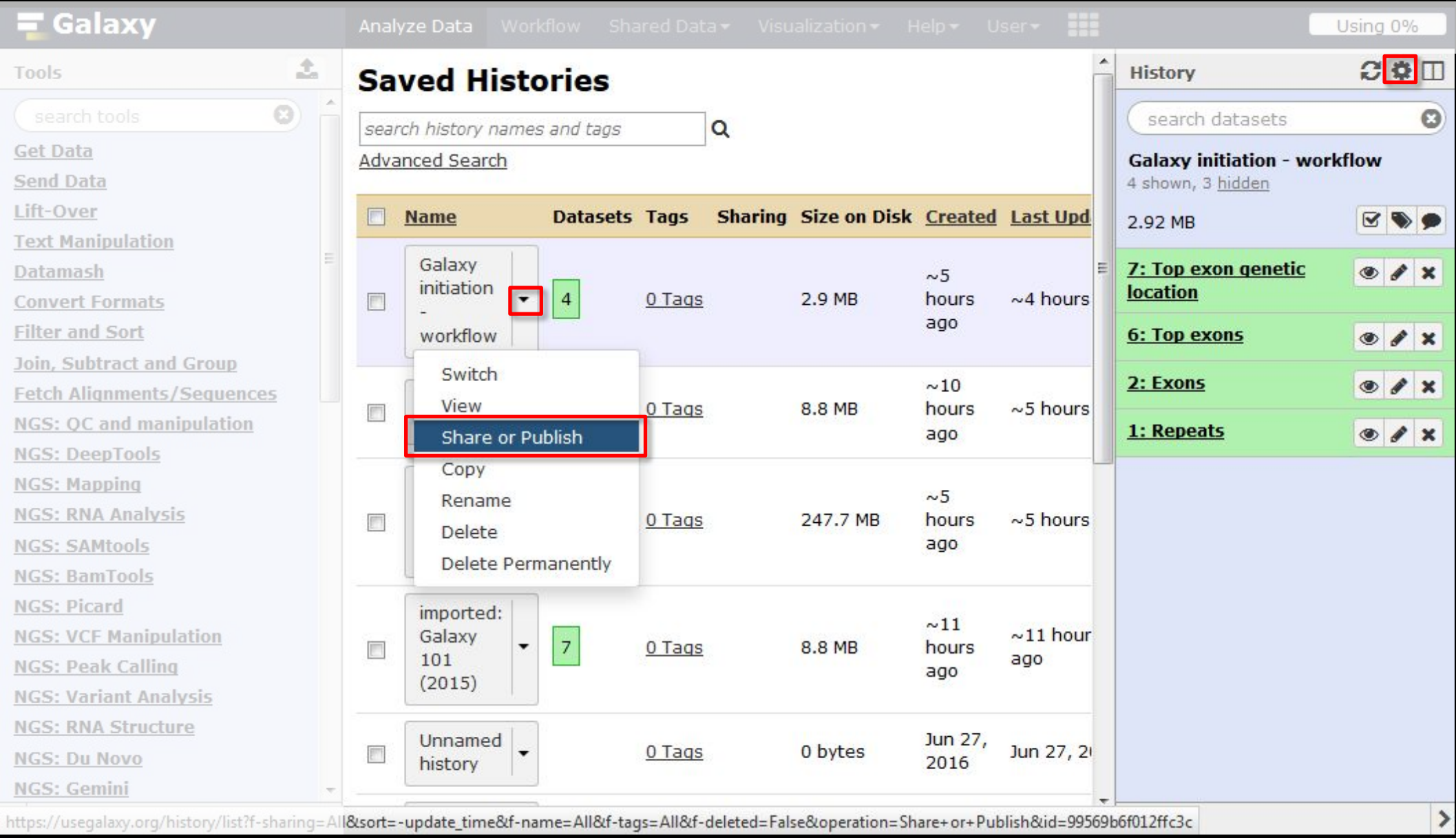
- Sharing workflows
  - Pre-configured parameters
  - With or without release parameters (set at runtime)
    - According to the user-end knowledge

bioinformatician ↔ bioinformatician

- Sharing tools ,scripts and wrappers
  - Toolshed

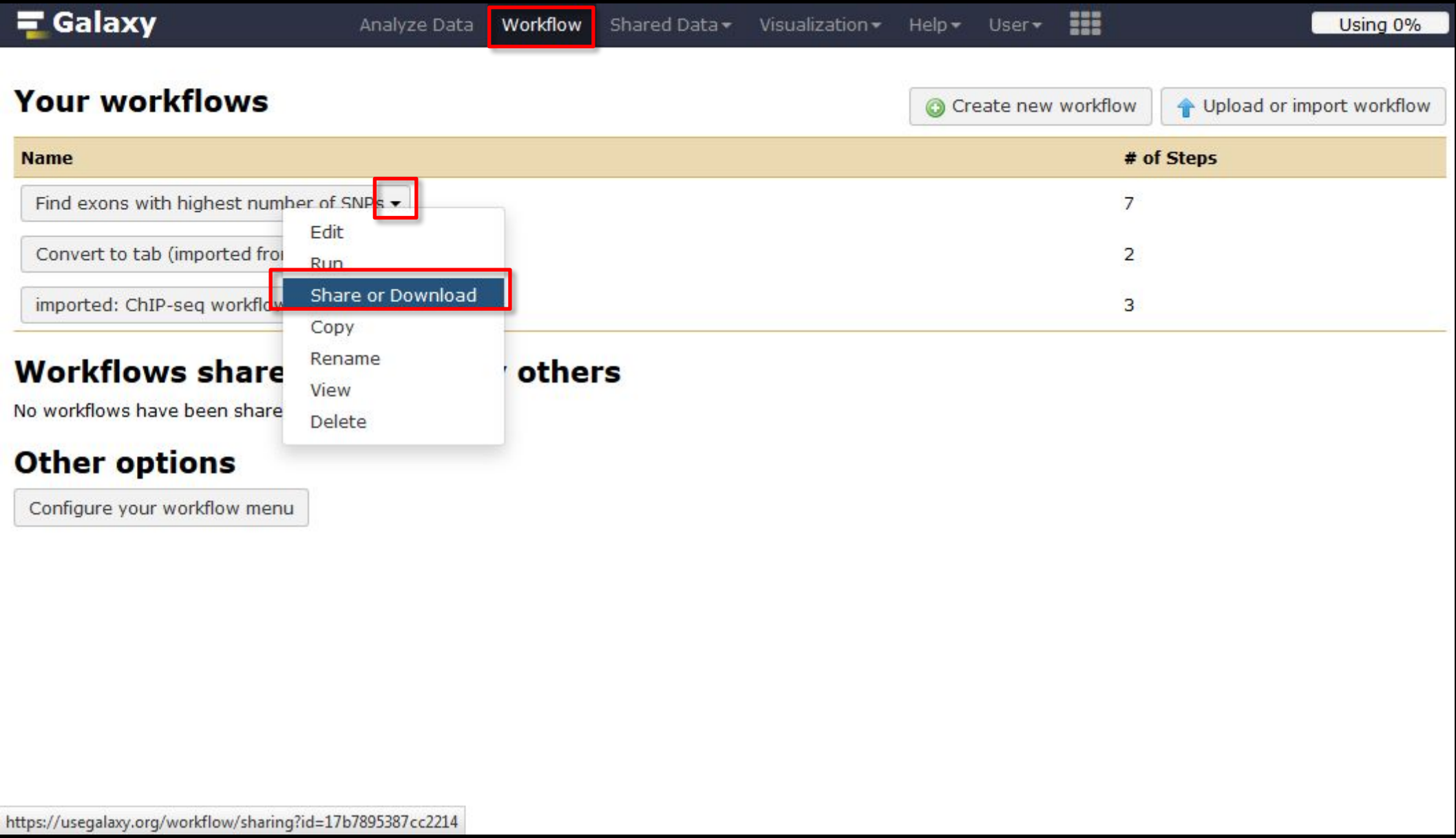
Share

History



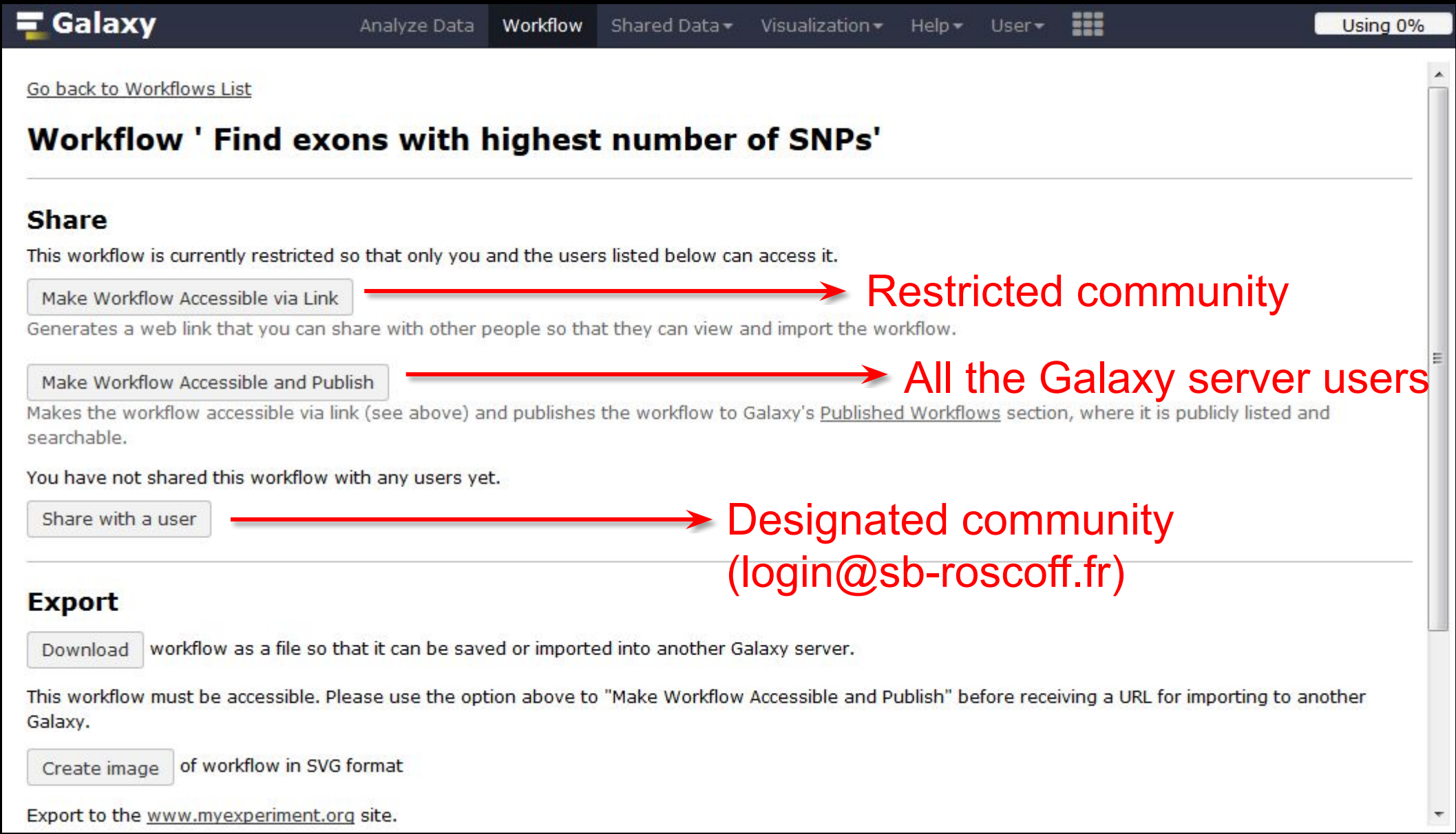
Share

Workflow

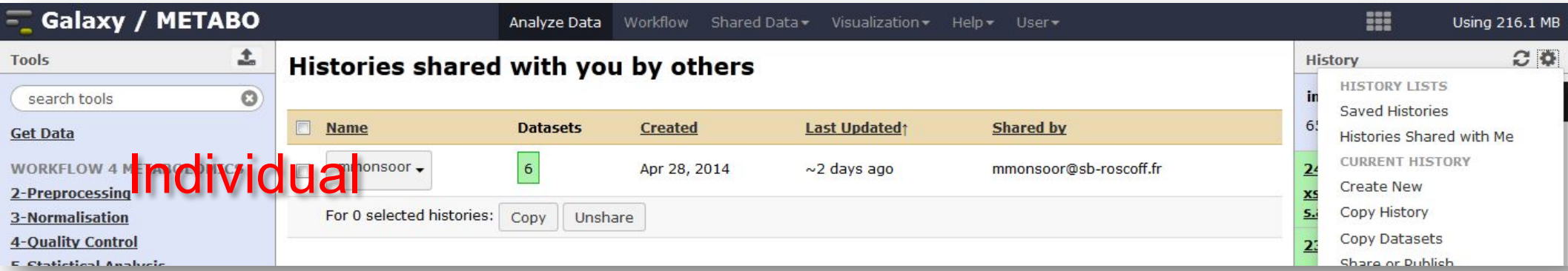


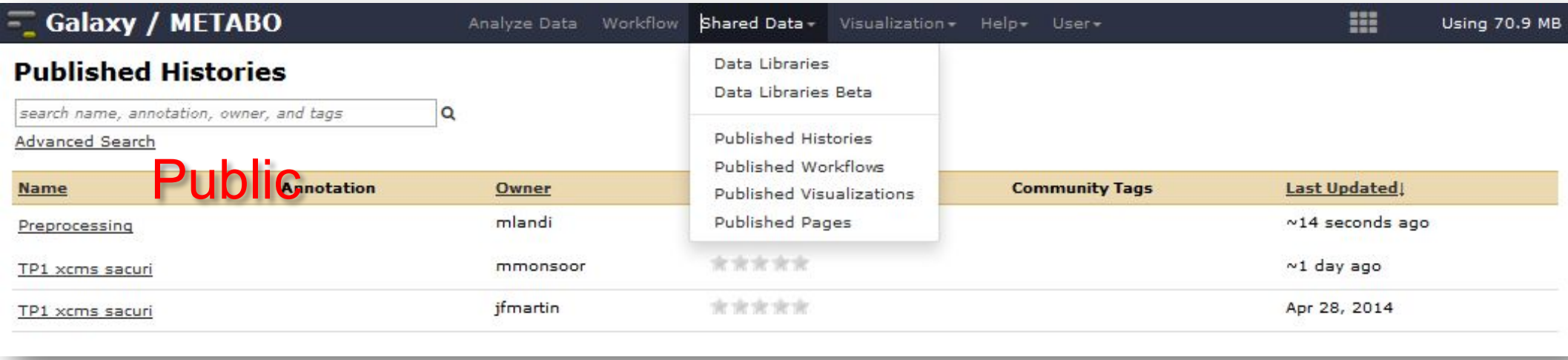
Share

Mode

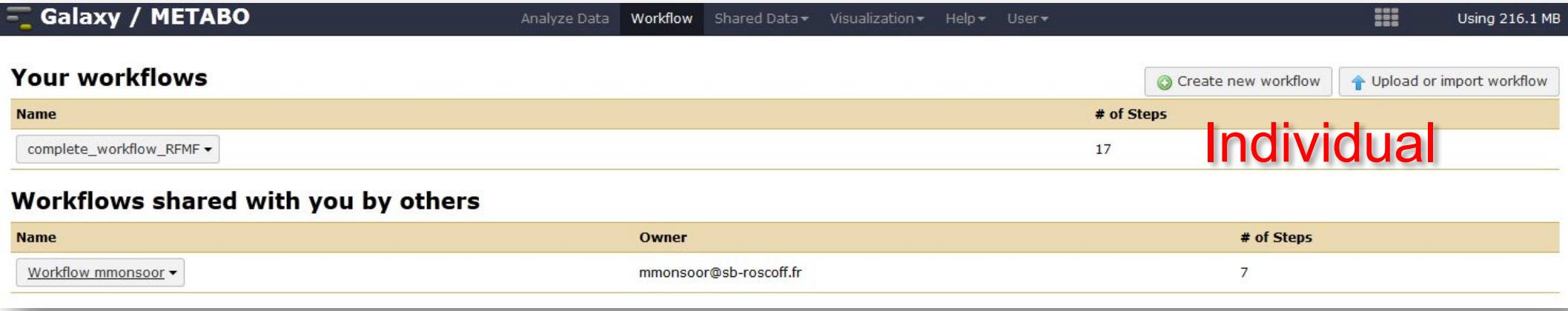

Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 0%
Go back to Workflows List
Workflow ' Find exons with highest number of SNPs'
Share
This workflow is currently restricted so that only you and the users listed below can access it.
Make Workflow Accessible via Link
Restricted community
Generates a web link that you can share with other people so that they can view and import the workflow.
Make Workflow Accessible and Publish
All the Galaxy server users
Makes the workflow accessible via link (see above) and publishes the workflow to Galaxy's Published Workflows section, where it is publicly listed and searchable.
You have not shared this workflow with any users yet.
Share with a user
Designated community
(login@sb-roscoff.fr)
Export
Download
workflow as a file so that it can be saved or imported into another Galaxy server.
This workflow must be accessible. Please use the option above to "Make Workflow Accessible and Publish" before receiving a URL for importing to another Galaxy.
Create image
of workflow in SVG format
Export to the www.myexperiment.org site.

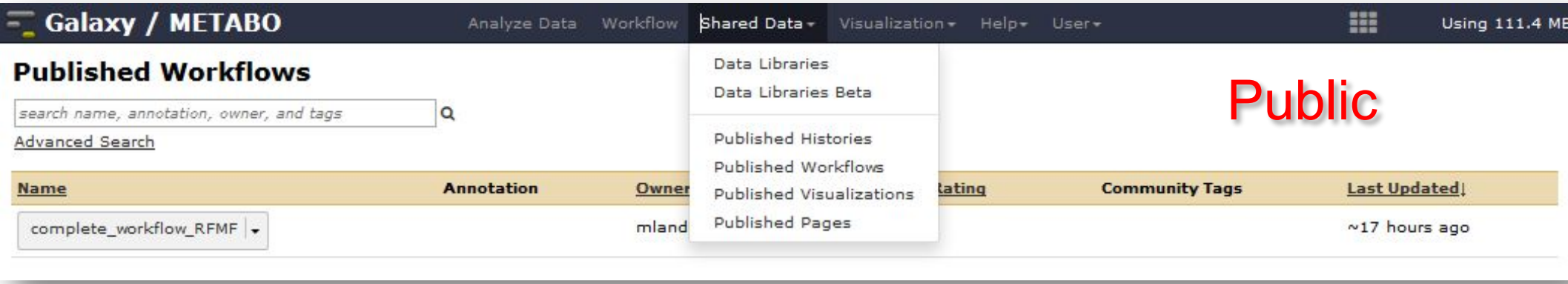
- Get shared histories



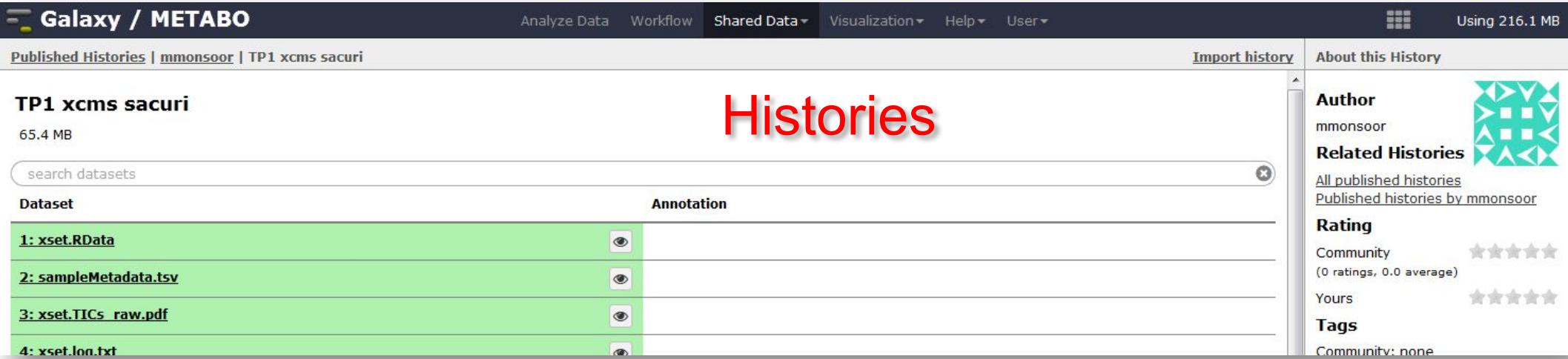


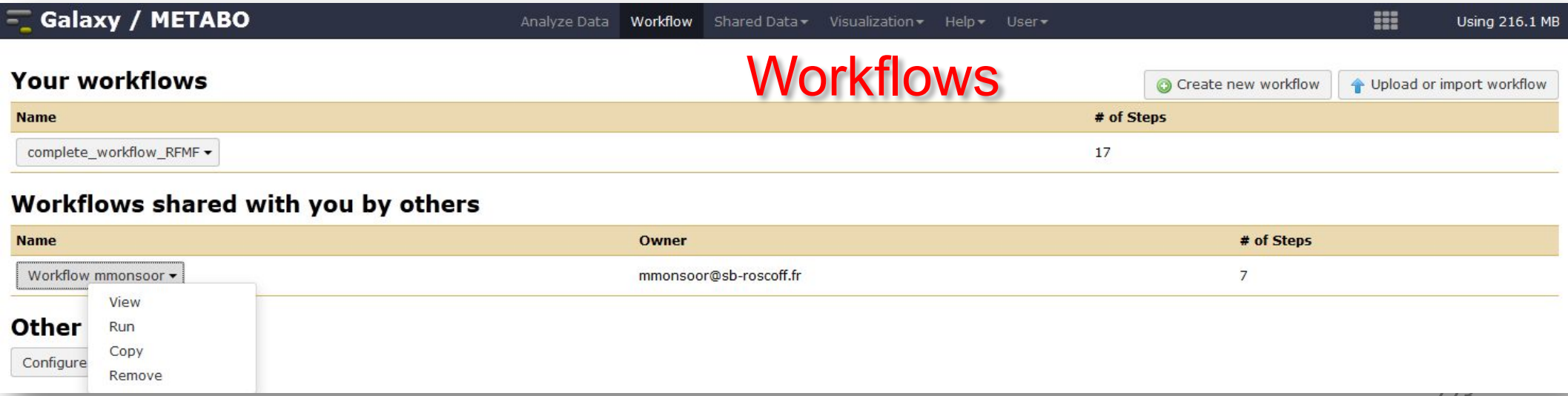
- Get shared workflows





- Import shared





223

Share

Level 5

- Share tools and descriptions in the ToolShed

Level 4

- Launch tools autonomously
- Use advanced parameters
- Use the Galaxy API
- Provide workflow for colleagues Level 1-3

Level 3

- Launch tools autonomously
- Use workflow more or less preset

Level 2

- Use preset workflow

Level 1

- Share his data to collegues Level 2-5

Hands-on

WORKFLOW

Workflow – Hands-on

Which coding exon has the highest number of ~~single nucleotide polymorphisms~~ repeats on chromosome 22?

1. Extract a workflow from your history
2. Edit the workflow (hide intermediate steps, rename inputs/outputs, set parameters at runtime, save)
3. Create a new history with the input data
4. Run the workflow
5. Share your history/workflow with your neighbour

CONCLUSION

Key points on Galaxy

- Easy-to-use graphical user interface for often complex command-line tools
- Keeps a full record of your analysis in a history
- Workflows enable you to repeat your analysis on different data
- Galaxy can connect to external sources for data import and visualization purposes
- Galaxy provides ways to share your results and methods with others

Tools - panel

125+ platforms for using Galaxy: https://galaxyproject.org/use
[Servers, clouds, deployable resources]

GENERAL PURPOSE:

usegalaxy.org, usegalaxy.eu, usegalaxy.fr, usegalaxy.org.au

DOMAIN SPECIFIC:

ABiMS

RNAseq: http://galaxy.sb-roscoff.fr
SBR tools: http://webtools.sb-roscoff.fr
Metagenomics: http://galaxy4metab.sb-roscoff.fr

Metabolomics: 4 Wm Workflow4metabolomics

Proteomics: ProteoRE

ChIP-seq: GalaxEast

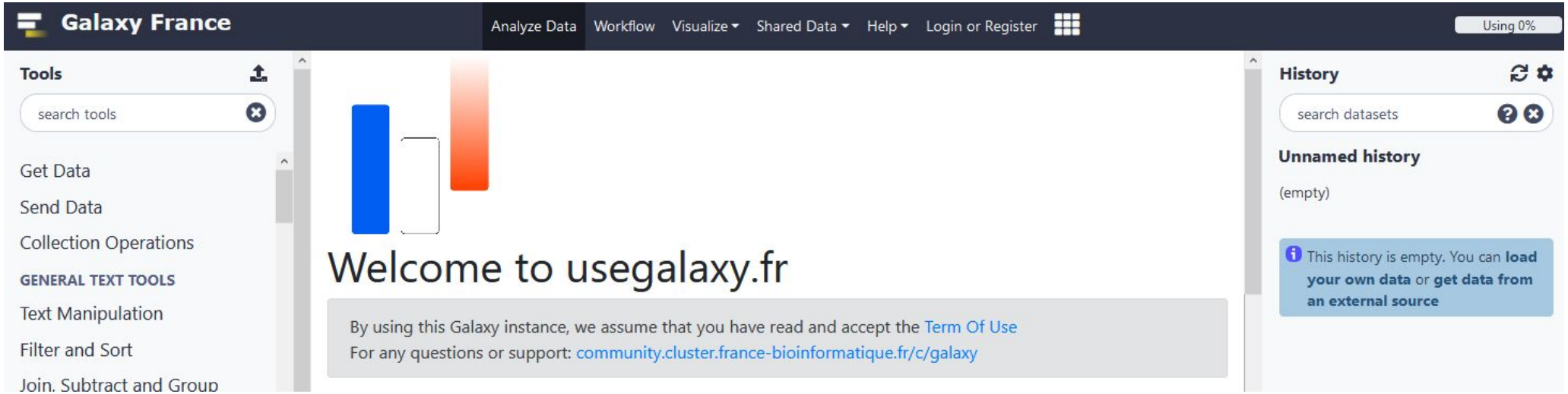
nebula ouris.fr

Catalog of French Galaxy tools:
http://galaxycat.france-bioinformatique.fr

Galaxy ABiMS --> Galaxy France

https://galaxy.sb-roscoff.fr/

https://usegalaxy.fr/



To go further

- On your own:
  - Training materials:

    https://galaxyproject.github.io/training-material
  - Interactive tours of Galaxy:

    http://galaxy.sb-roscoff.fr/tours

- Training courses:

Training	What ?	Where ?	When ?
RNAseq analysis with Galaxy	RNAseq	Roscoff, France	first semester 2021
Galaxy Community Conference (GCC)	General purpose (data-intensive biology and Galaxy)	Ghent, Belgium	July 2021
Workflow4Experimenters	Metabolomics	Toulouse, France	October 2021

END

Training evaluation

Thank you for completing the training evaluation questionnaire:

http://abims.sb-roscoff.fr/evaluation_formation

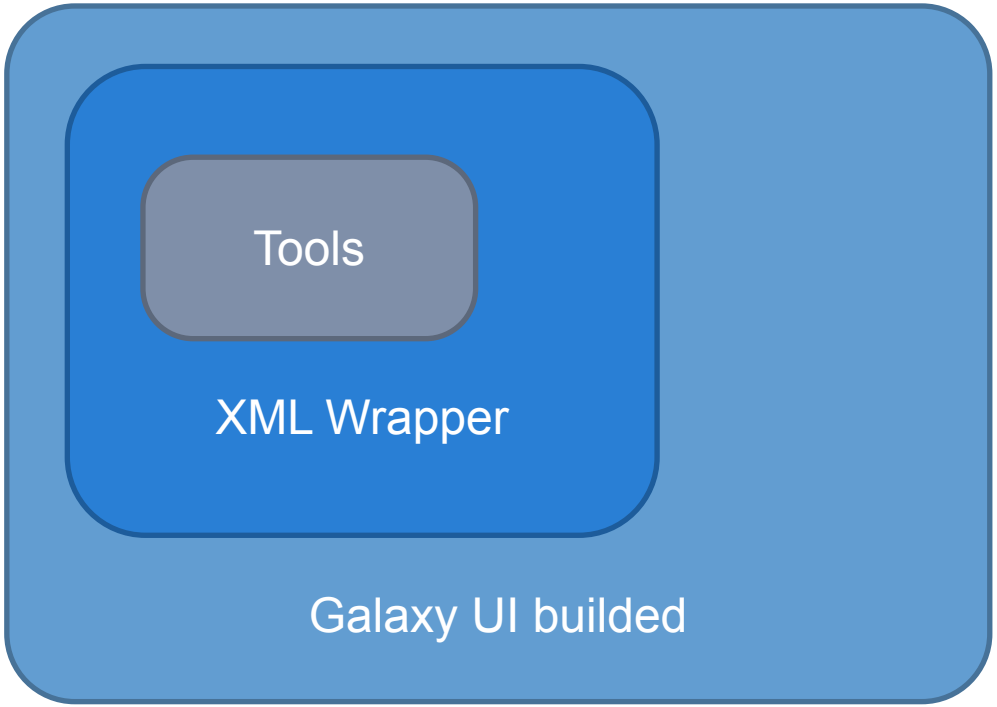
BONUS

How are tools born?

BONUS

Bonus

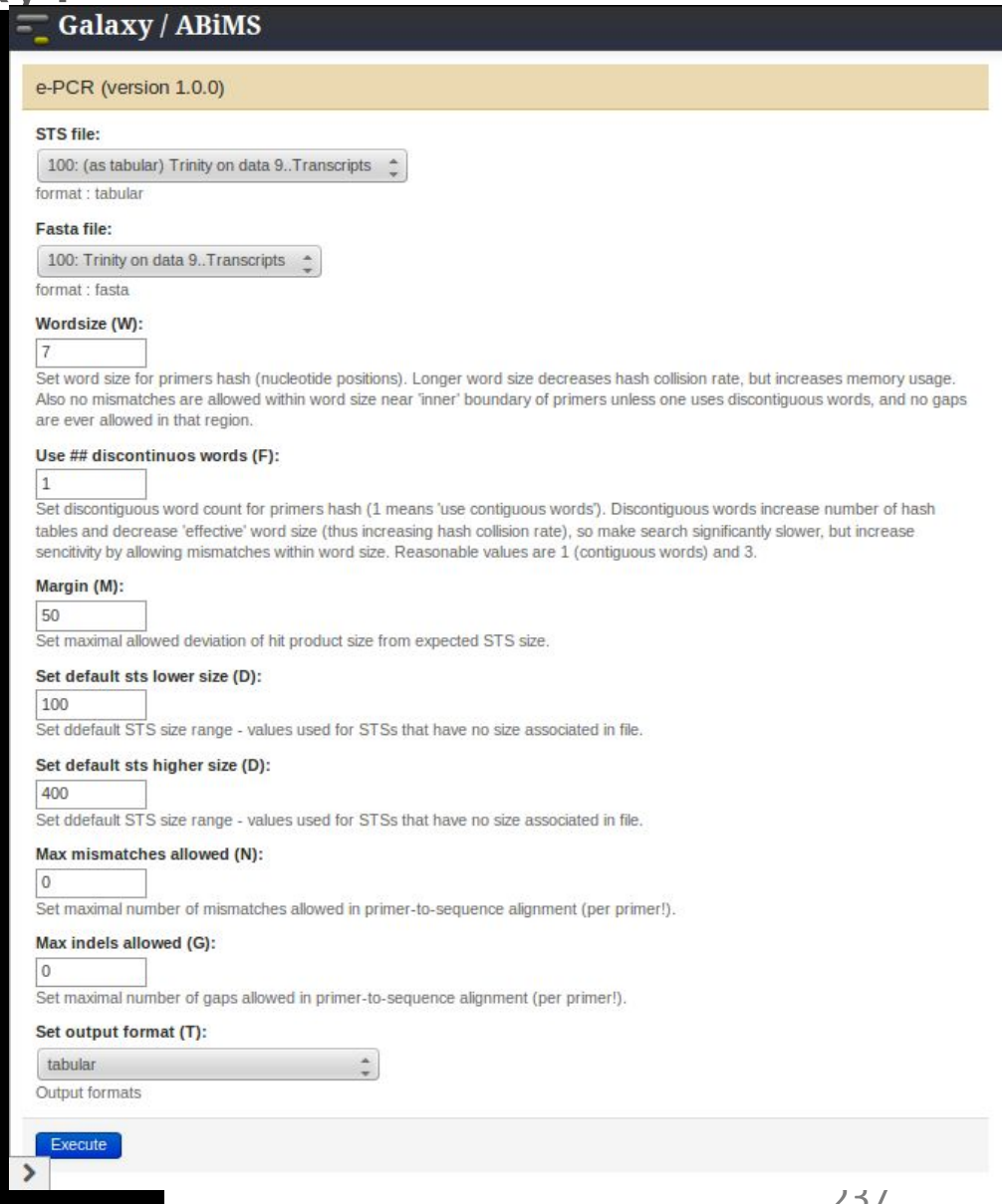
- How to import a tool in Galaxy?



Bonus

- How to import a tool in Galaxy?

```
[lecorguille@n0 ~]$ e-PCR --help
e-PCR: invalid option -- -
usage: [-hV] [posix-options] stsfile [fasta ...]
[compat-options]
where posix-options are:
    -m ##    Margin (default 50)
    -w ##    Wordsize  (default 7)
    -n ##    Max mismatches allowed (default 0)
    -g ##    Max indels allowed (default 0)
    -f ##    Use ## discontiguos words, slow if
              ##>1
    -o ##    Set output file
    -t ##    Set output format:
         1 - classic, range (pos1..pos2)
         2 - classic, midpoint
         3 - tabular
         4 - tabular with alignment in comments
              (slow)
    -d##-##  Set default size range
              (default 100-350)
    -p +-    Turn hits postprocess on/off
    -v ##    Verbosity flags
    -a a|f   Use presize alignmens (only if
              gaps>0), slow
          a - Allways or f - as Fallback
    -x +-    Use 5'-end lowercase masking of
              primers (default -)
    -u +-    Uppercase all primers (default -)
```

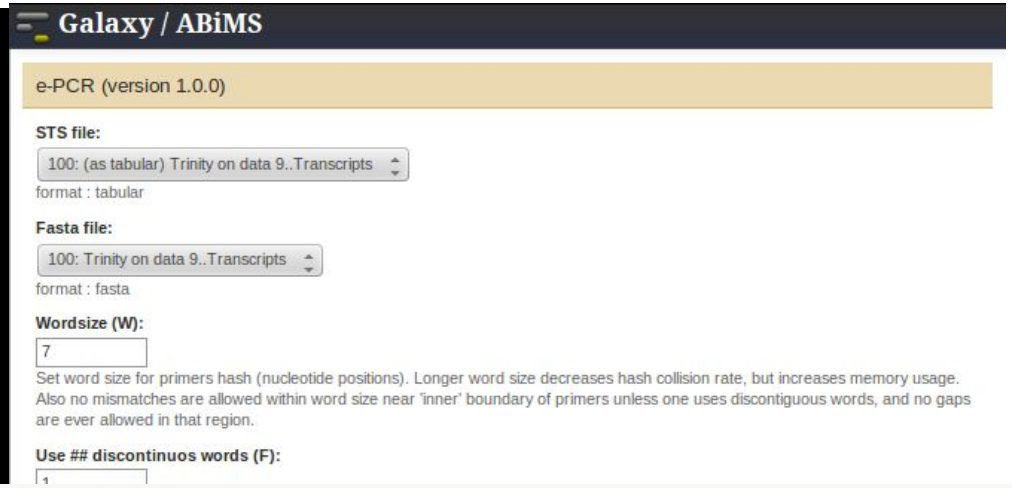


- How to import a tool in Galaxy?

```
<tool id="abims_epcr" name="e-PCR">
    <!-- author : lecorguille@sb-roscoff.fr -->
    <!-- date : 11-05-12 -->
    <description>e-PCR parses stsfile in unists format, then reads nucleotide sequence data in FASTA format from files listed in commandline if
    any, or from stdin otherwise. For input sequences e-PCR finds matches and prints output in one of three formats.</description>

    <command>e-PCR -w $wordsize -f $wordcnt -m $margin -d$sts_size_lo-$sts_size_hi -n $max_mismatch -g $max_gap -t $output_format $infile_stsfile
    $infile_fasta > $output</command>

    <inputs>
        <param name="infile_stsfile" type="data" label="STS file" format="tabular" help="format : tabular" />
        <param name="infile_fasta" type="data" label="Fasta file" format="fasta" help="format : fasta" />
        <param name="wordsize" type="integer" label="Wordsize (W)" value="7" help="Set word size for primers hash (nucleotide positions).
        Longer word size decreases hash collision rate, but increases memory usage. Also no mismatches are allowed within word size near
        'inner' boundary of primers unless one uses discontiguous words, and no gaps are ever allowed in that region." />
        <param name="wordcnt" type="integer" label="Use ## discontinuos words (F)" value="1" help="Set discontiguous word count for primers
        hash (1 means 'use contiguous words'). Discontiguous words increase number of hash tables and decrease 'effective' word size (thus
        increasing hash collision rate), so make search significantly slower, but increase sencitivity by allowing mismatches within word
        size. Reasonable values are 1 (contiguous words) and 3." />
        <param name="margin" type="integer" label="Margin (M)" value="50" help="Set maximal allowed deviation of hit product size from
        expected STS size." />
        <param name="sts_size_lo" type="integer" label="Set default sts lower size (D)" value="100" help="Set ddefault STS size range - values
        used for STSs that have no size associated in file." />
        <param name="sts_size_hi" type="integer" label="Set default sts higher size (D)" value="400" help="Set ddefault STS size range -
        values used for STSs that have no size associated in file." />
        <param name="max_mismatch" type="integer" label="Max mismatches allowed (N)" value="0" help="Set maximal number of mismatches allowed
        in primer-to-sequence alignment (per primer!)." />
        <param name="max_gap" type="integer" label="Max indels allowed (G)" value="0" help="Set maximal number of gaps allowed in primer-to-
        sequence alignment (per primer!)." />
        <param name="output_format" type="select" help="Output formats">
            <label>Set output format (T)</label>
            <option value="1">classic, range (pos1..pos2)</option>
            <option value="2">classic, midpoint</option>
            <option value="3" selected="true">tabular</option>
            <option value="4">tabular with alignment in comments (slow)</option>
        </param>
    </inputs>

    <outputs>
        <data name="output" format="tabular" />
    </outputs>
```

- How to import a tool in Galaxy?

```
[lecorguille@n0 ~]$ e-PCR --help
e-PCR: invalid option -- -
usage: [-hV] [posix-options] stsfile [fasta ...]
[compat-options]
where posix-options are:
     -m ##     Margin (default 50)
     -w ##     Wordsize  (default 7)
     -n ##     Max mismatches allowed (default 0)
     -g ##     Max indels allowed (default 0)
     -f ##     Use ## discontiguos words, slow if
                ##>1
```




```
<tool id="abims_epcr" name="e-PCR">
      <!-- author : lecorguille@sb-roscoff.fr -->
      <!-- date : 11-05-12 -->
      <description>e-PCR parses stsfile in unists format, then reads nucleotide sequence data in FASTA format from files listed in commandline if
      any, or from stdin otherwise. For input sequences e-PCR finds matches and prints output in one of three formats.</description>
      
      <command>e-PCR -w $wordsize -f $wordcnt -m $margin -d$sts_size_lo-$sts_size_hi -n $max_mismatch -g $max_gap -t $output_format $infile_stsfile
      $infile_fasta > $output</command>

      <inputs>
            <param name="infile_stsfile" type="data" label="STS file" format="tabular" help="format : tabular" />
            <param name="infile_fasta" type="data" label="Fasta file" format="fasta" help="format : fasta" />
            <param name="wordsize" type="integer" label="Wordsize (W)" value="7" help="Set word size for primers hash (nucleotide positions).
            Longer word size decreases hash collision rate, but increases memory usage. Also no mismatches are allowed within word size near
            'inner' boundary of primers unless one uses discontiguous words, and no gaps are ever allowed in that region." />
            <param name="wordcnt" type="integer" label="Use ## discontinuos words (F)" value="1" help="Set discontiguous word count for primers
            hash (1 means 'use contiguous words'). Discontiguous words increase number of hash tables and decrease 'effective' word size (thus
            increasing hash collision rate), so make search significantly slower, but increase sencitivity by allowing mismatches within word
            size. Reasonable values are 1 (contiguous words) and 3." />
            <param name="margin" type="integer" label="Margin (M)" value="50" help="Set maximal allowed deviation of hit product size from
            expected STS size." />
            <param name="sts_size_lo" type="integer" label="Set default sts lower size (D)" value="100" help="Set ddefault STS size range - values
            used for STSs that have no size associated in file." />
            <param name="sts_size_hi" type="integer" label="Set default sts higher size (D)" value="400" help="Set ddefault STS size range -
            values used for STSs that have no size associated in file." />
            <param name="max_mismatch" type="integer" label="Max mismatches allowed (N)" value="0" help="Set maximal number of mismatches allowed
```