oqtans

Quantitative Transcriptome Analysis in the Cloud — Part I

Géraldine Jean <Geraldine.Jean@univ-nantes.fr>
Department of Computer Science, Combinatorics and Bioinformatics Group, LINA — University of Nantes, France
Sebastian J. Schultheiss <sebastian@bioweb.me>
Computomics Molecular Data Analysis, Center for Bioinformatics, University of Tübingen, Germany

Presented at the Meeting on Advances and Challenges of RNA-Seq Analysis Halle, Germany

UNIVERSITÉ DE NANTES

MAX-PLANCK-GESELLSCHAFT

computomics®
molecular data analysis

oqtans
online quantitative transcript analysis

Galaxy

Galaxy Workbench

- Developed at Penn State and Emory
- Open, web-based platform
- Data intensive biomedical research
- Accessible: public web server; download and install; cloud deployment

J. Goecks et al. 2010
D. Blankenberg et al. 2010
E. Afgan et al. 2010
S. Koskovsky Pond et al. 2009

W. Miller et al. 2007
J. Taylor et al. 2007
D. Blankenberg et al. 2007
M. Giardine et al. 2005

Galaxy Approach

- Persistent (histories, parameters, versions)
- Reproducible (histories, workflows, pages)
- Transparent (accessible, easy to use)
- Credibility for bioinformatics research

J. Goecks et al. 2010
D. Blankenberg et al. 2010
E. Afgan et al. 2010
S. Koskovsky Pond et al. 2009

W. Miller et al. 2007
J. Taylor et al. 2007
D. Blankenberg et al. 2007
M. Giardine et al. 2005

Galaxy Approach

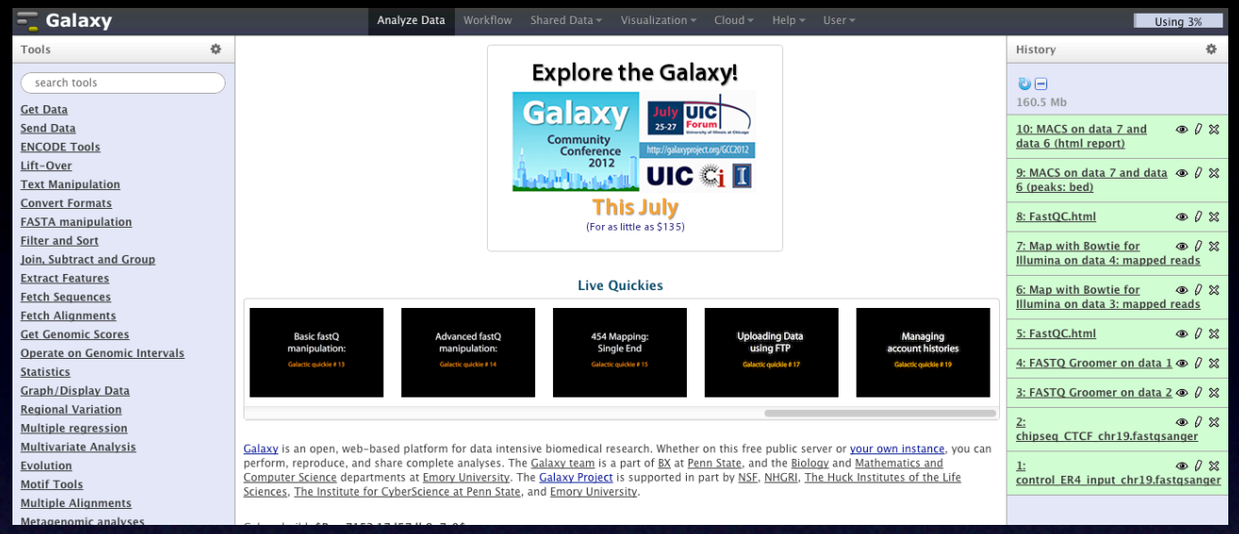
- Bioinformaticians: Integration made simple with xml wrappers of command line tools
- Biologists: Easy to use interface in any web browser
- You: Large collection of RNA-seq tools

J. Goecks et al. 2010
D. Blankenberg et al. 2010
E. Afgan et al. 2010
S. Koskovsky Pond et al. 2009

W. Miller et al. 2007
J. Taylor et al. 2007
D. Blankenberg et al. 2007
M. Giardine et al. 2005

Methods of Access

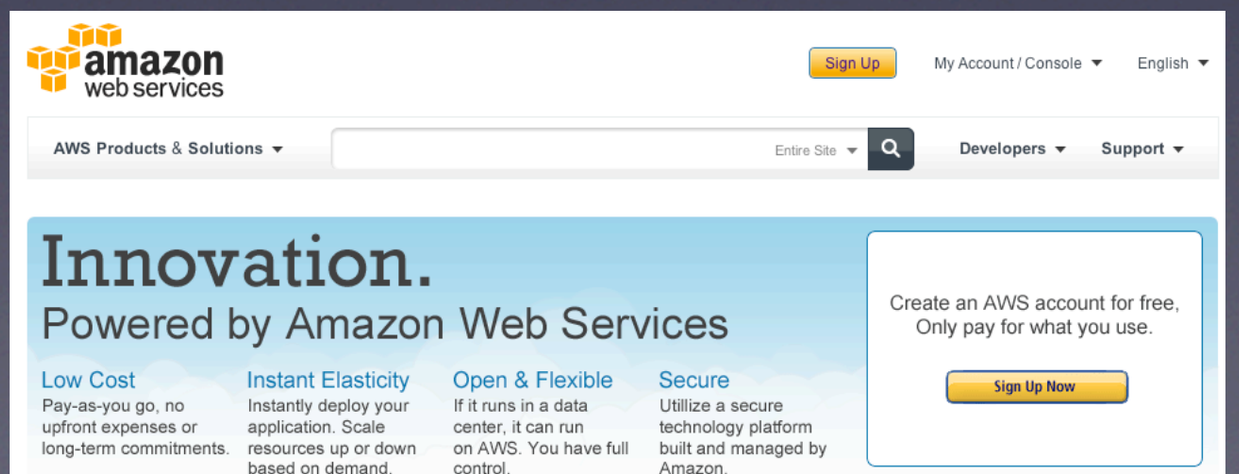
- http://usegalaxy.org



- http://getgalaxy.org



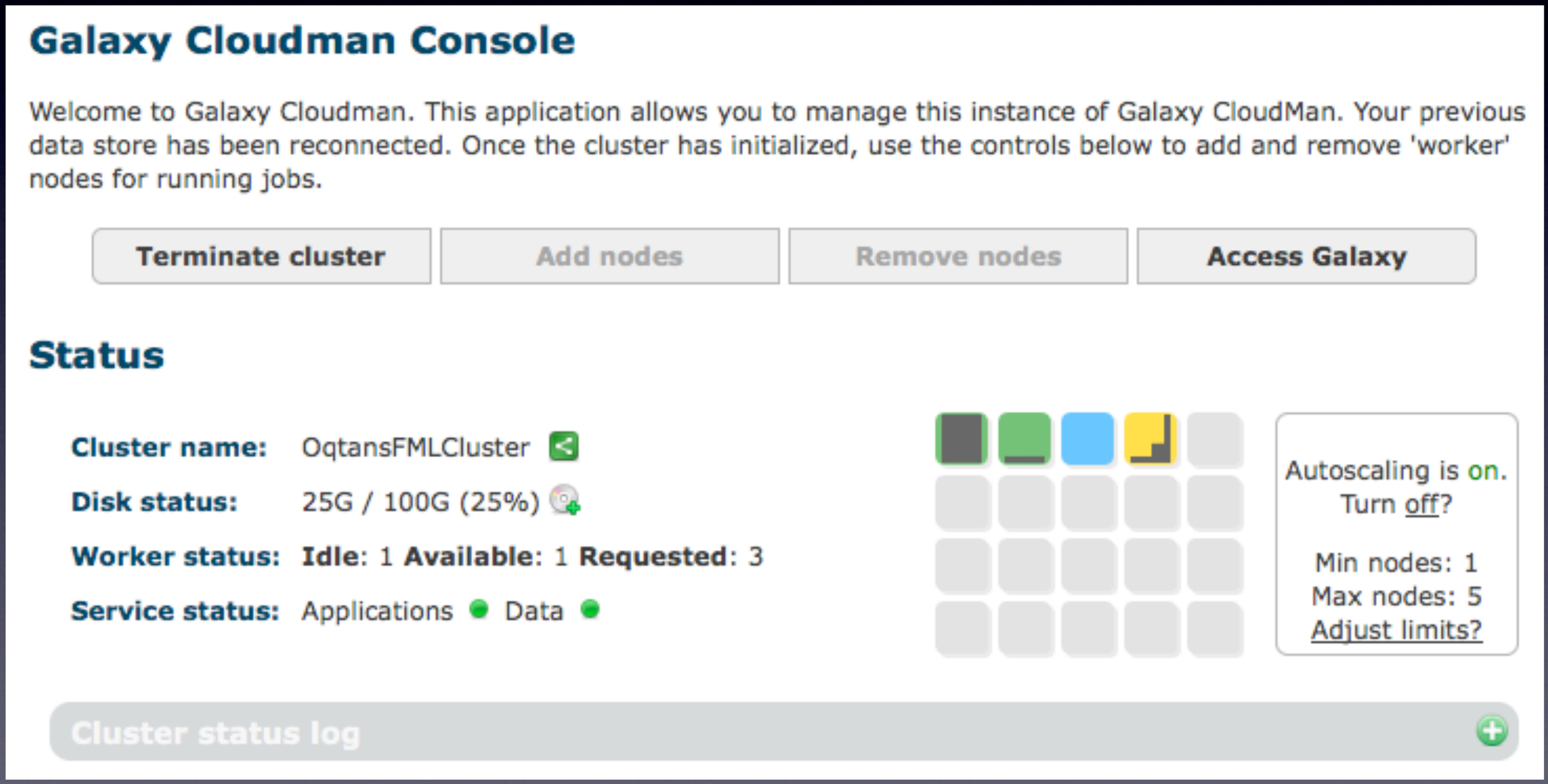
- aws.amazon.com



Methods of Access

- http://bioweb.me/mlb-galaxy
- http://cloud.oqtans.org
- http://oqtans.org

Cloud Computing



Galaxy Pages

- Describe your work to others
- Create the “supplementary materials” page
- Annotate history items and workflows
- Make a tutorial for your tools

Galaxy Pages

Galaxy 101 Tutorial with *Arabidopsis thaliana* data

This example is based on Anton Nekrutenko's Galaxy 101 Tutorial and repeats these steps using data from *A. thaliana*.

We're comparing the position of SNPs with the genome annotation of Chromosome 1 of Columbia-0. For this, we need two .bed files with the data:

Galaxy Dataset | col-0_snp.bed

Galaxy Dataset | col-0_exons.bed

We combine these two datasets in the following workflow and count the number of SNPs per exon:

Galaxy Workflow | Exons and SNPs (Galaxy 101 Example)

Step	Annotation
Step 1: Input dataset	
Exons *select at runtime*	
Step 2: Input dataset	
SNPs *select at runtime*	

Author
sebastian-schultheiss

Related Pages

All published pages
Published pages by sebastian-schultheiss

Rating

Community
(1 ratings, 5 average)

Yours
(thanks!)

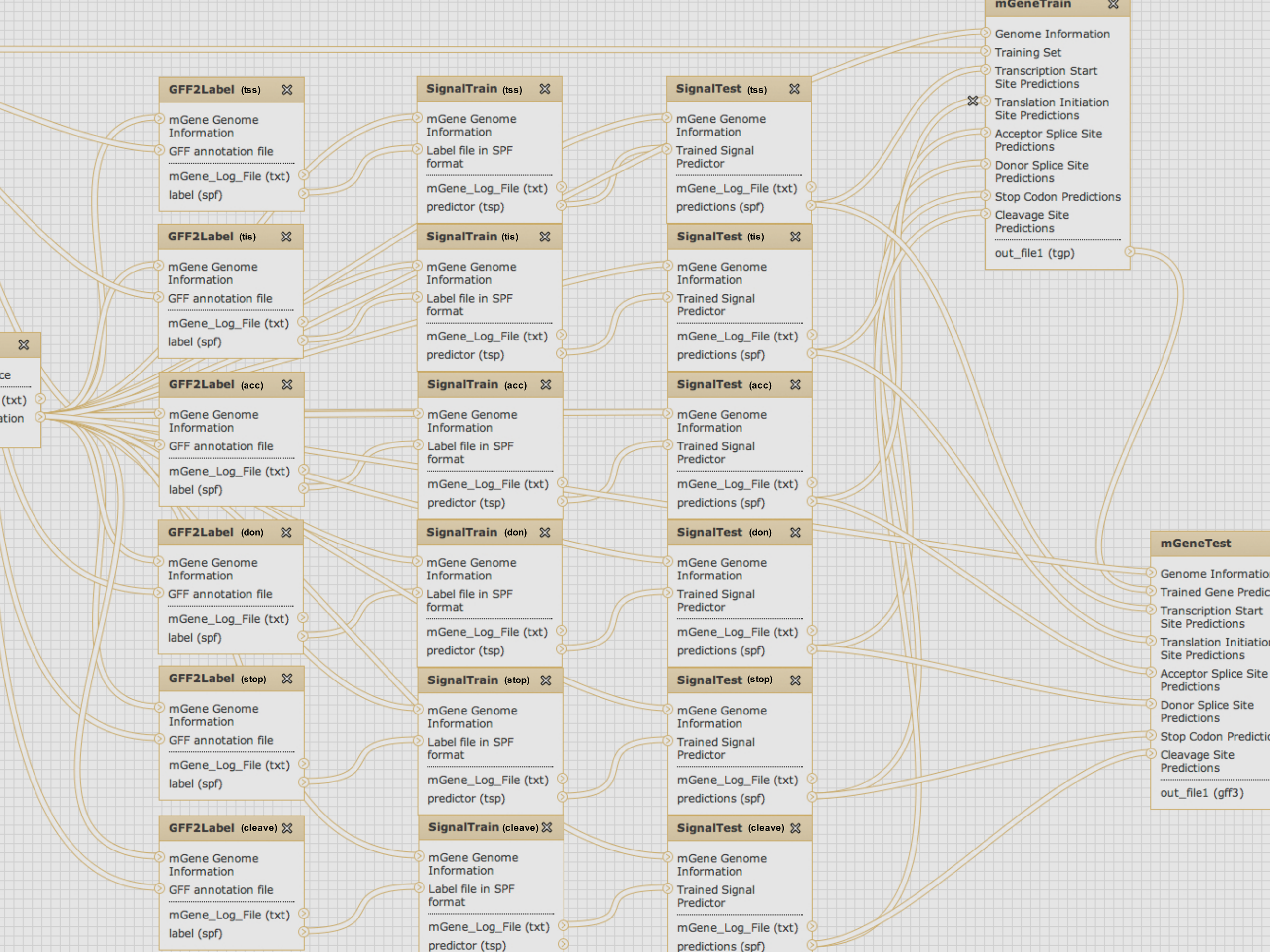
Tags

Community: none

Yours:

Galaxy Workflows

- Automate common tasks
- Create from scratch or from existing history
- Fill in all parameters at the beginning
- Share workflows with others


GFF2Label (tss)
mGene Genome Information
GFF annotation file
mGene_Log_File (txt)
label (spf)
SignalTrain (tss)
mGene Genome Information
Label file in SPF format
mGene_Log_File (txt)
predictor (tsp)
SignalTest (tss)
mGene Genome Information
Trained Signal Predictor
mGene_Log_File (txt)
predictions (spf)
mGeneTrain
Genome Information
Training Set
Transcription Start Site Predictions
Translation Initiation Site Predictions
Acceptor Splice Site Predictions
Donor Splice Site Predictions
Stop Codon Predictions
Cleavage Site Predictions
out_file1 (tgp)
GFF2Label (tis)
SignalTrain (tis)
SignalTest (tis)
GFF2Label (acc)
SignalTrain (acc)
SignalTest (acc)
GFF2Label (don)
SignalTrain (don)
SignalTest (don)
GFF2Label (stop)
SignalTrain (stop)
SignalTest (stop)
GFF2Label (cleave)
SignalTrain (cleave)
SignalTest (cleave)
mGeneTest
Genome Information
Trained Gene Predic
Transcription Start Site Predictions
Translation Initiation Site Predictions
Acceptor Splice Site Predictions
Donor Splice Site Predictions
Stop Codon Predictio
Cleavage Site Predictions
out_file1 (gff3)

Galaxy 101
Live Demonstration

- Show you how to perform an experiment
- Obtain data from external sources
- Perform data manipulation and display data
- Understand Galaxy's History system
- Create, modify, apply workflows to data

based on A. Nekrutenko et al. (2010) Galaxy 101 Tutorial
https://main.g2.bx.psu.edu/galaxy101

Galaxy
Live Demonstration



- Coding Exons
- SNPs

- Join Exons & SNPs

- Count SNPs/ Exon
- Sort

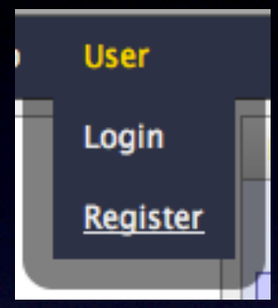
- Select top 5
- Display

Get Data

- Create a Galaxy account
- Getting data from a library

Get Data

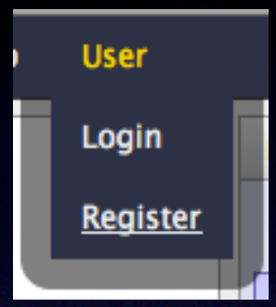
- Create a Galaxy account



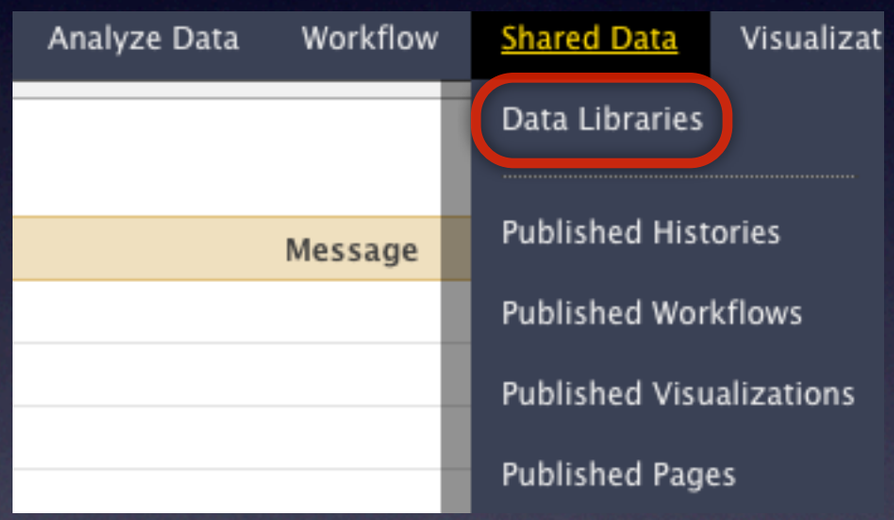
- Getting data from a library

Get Data

- Create a Galaxy account

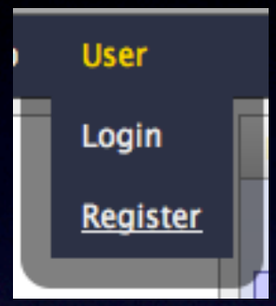


- Getting data from a library

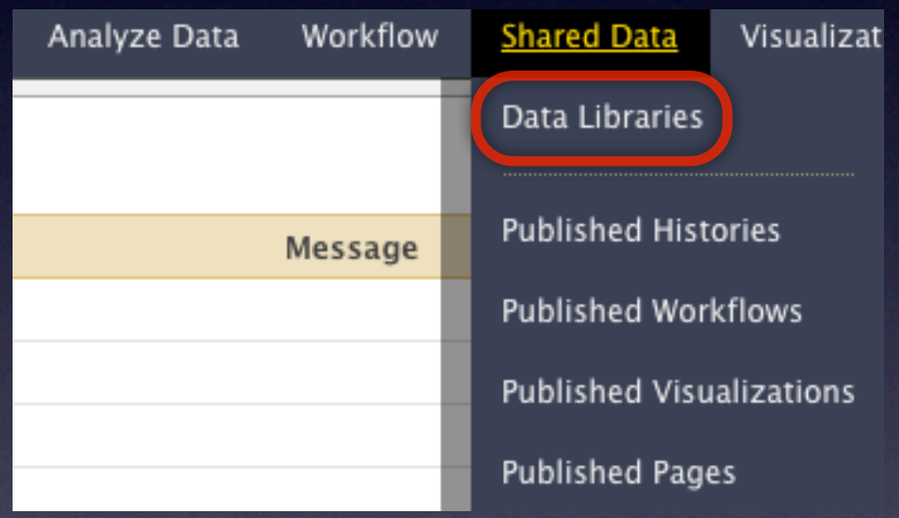


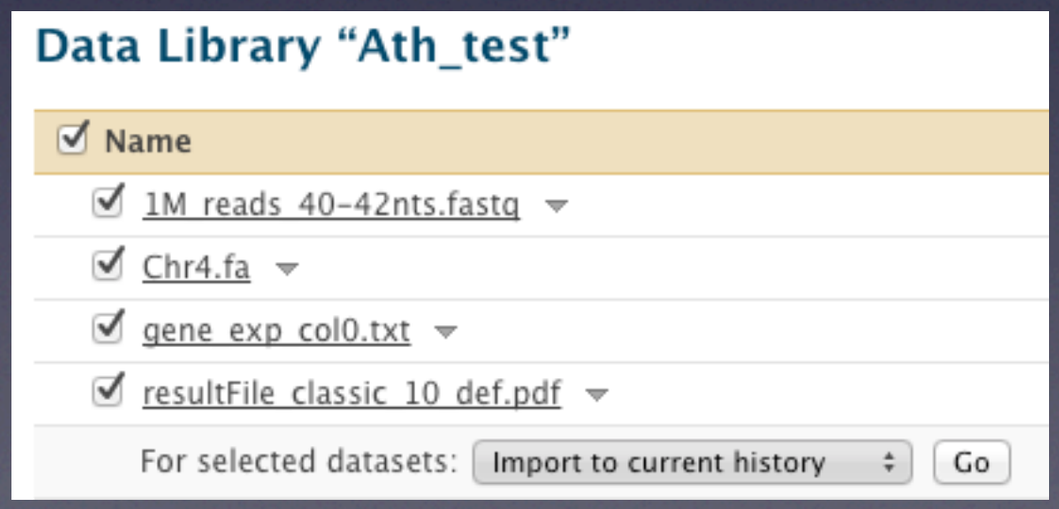
Get Data

- Create a Galaxy account



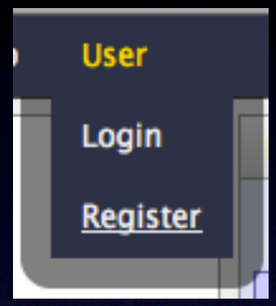
- Getting data from a library



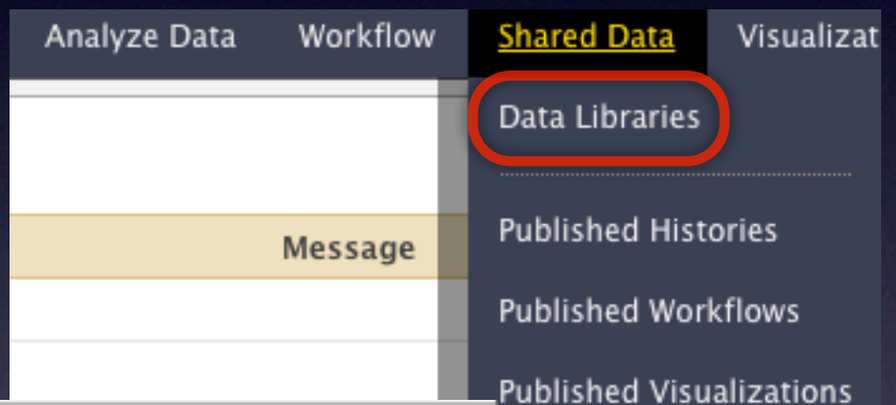


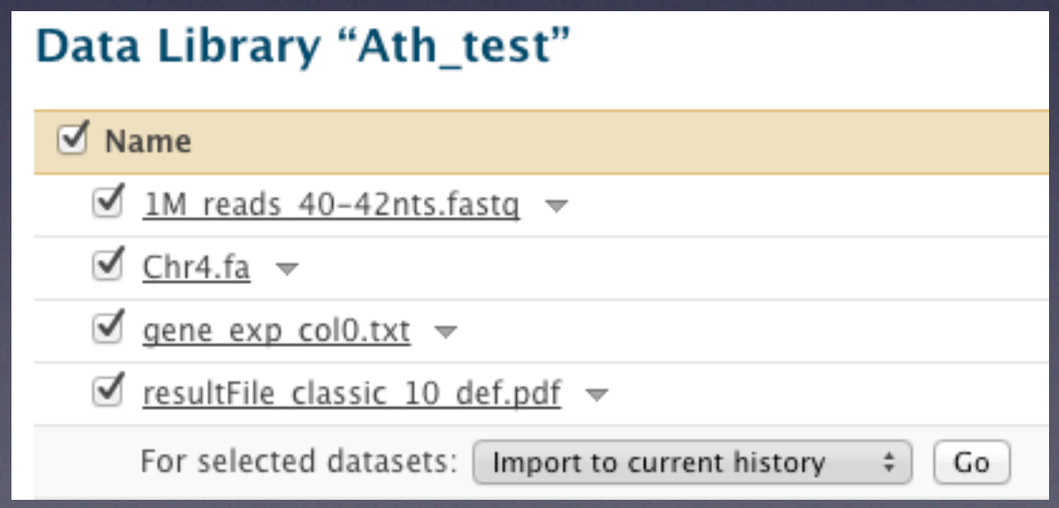
Get Data

- Create a Galaxy account



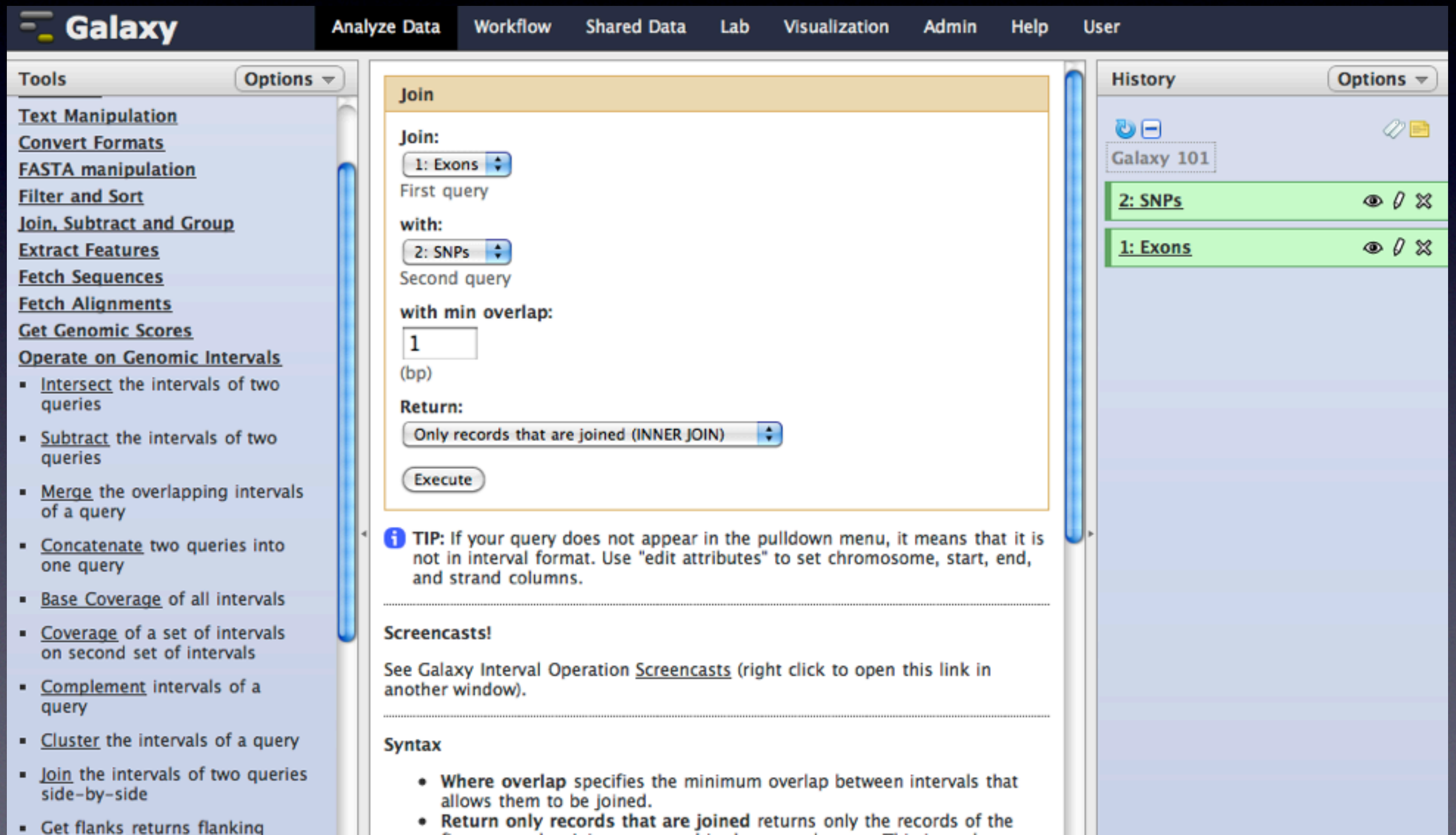
- Getting data from a library



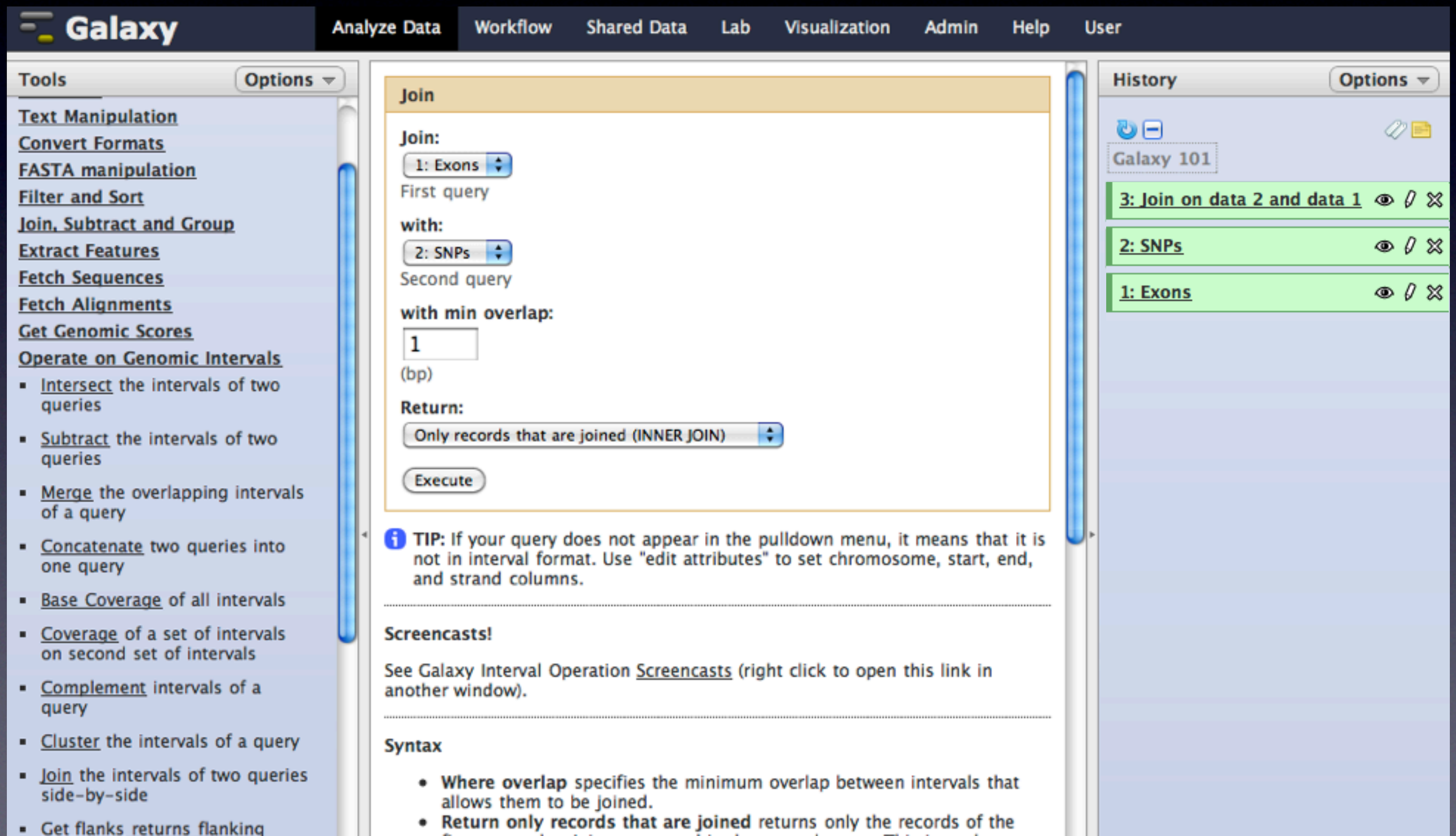




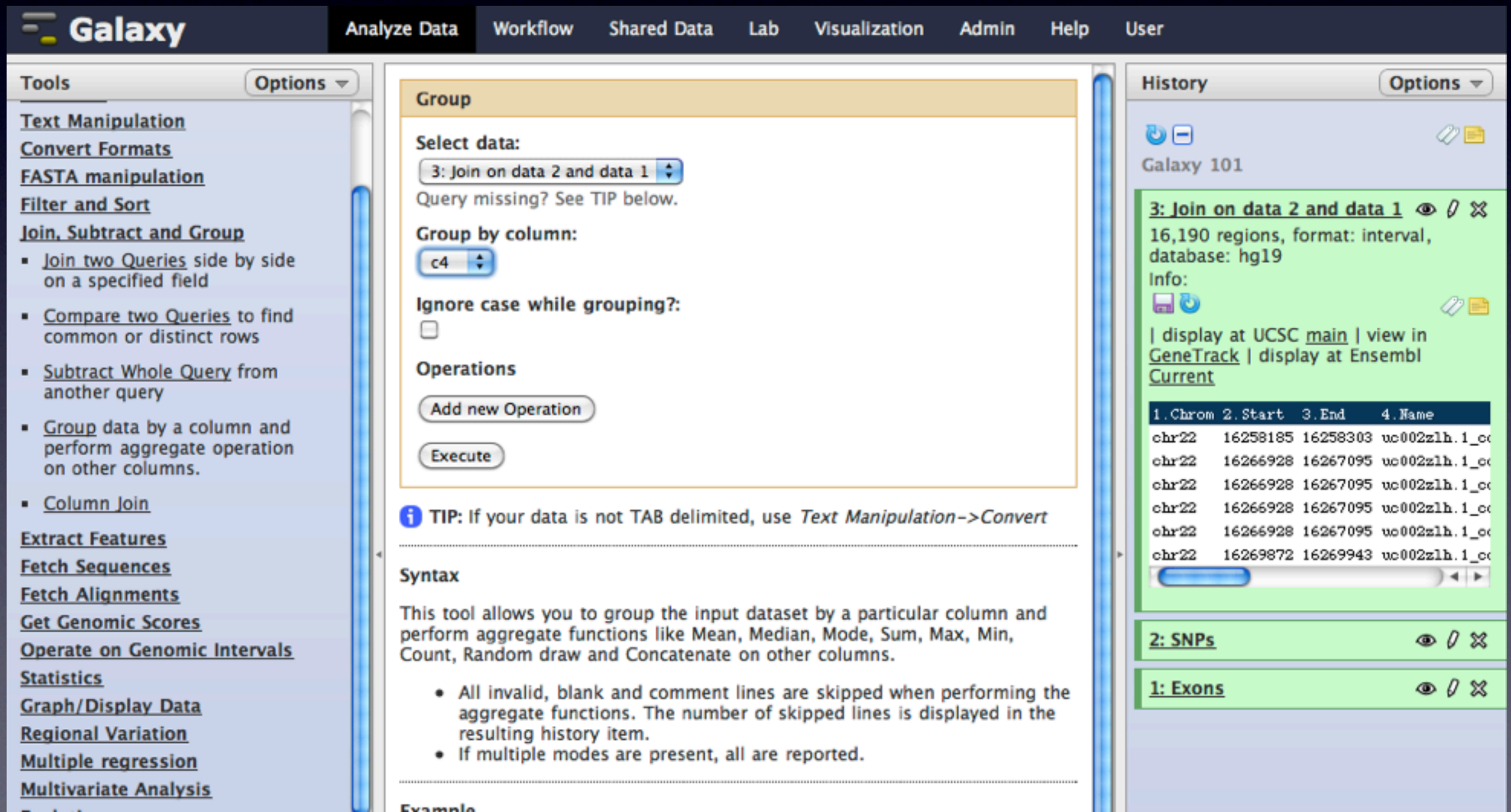
Join Data


Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
Tools
Options
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Intersect the intervals of two queries
Subtract the intervals of two queries
Merge the overlapping intervals of a query
Concatenate two queries into one query
Base Coverage of all intervals
Coverage of a set of intervals on second set of intervals
Complement intervals of a query
Cluster the intervals of a query
Join the intervals of two queries side-by-side
Get flanks returns flanking
Join
Join:
1: Exons
First query
with:
2: SNPs
Second query
with min overlap:
1
(bp)
Return:
Only records that are joined (INNER JOIN)
Execute
TIP: If your query does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.
Screencasts!
See Galaxy Interval Operation Screencasts (right click to open this link in another window).
Syntax
Where overlap specifies the minimum overlap between intervals that allows them to be joined.
Return only records that are joined returns only the records of the
History
Options
Galaxy 101
2: SNPs
1: Exons

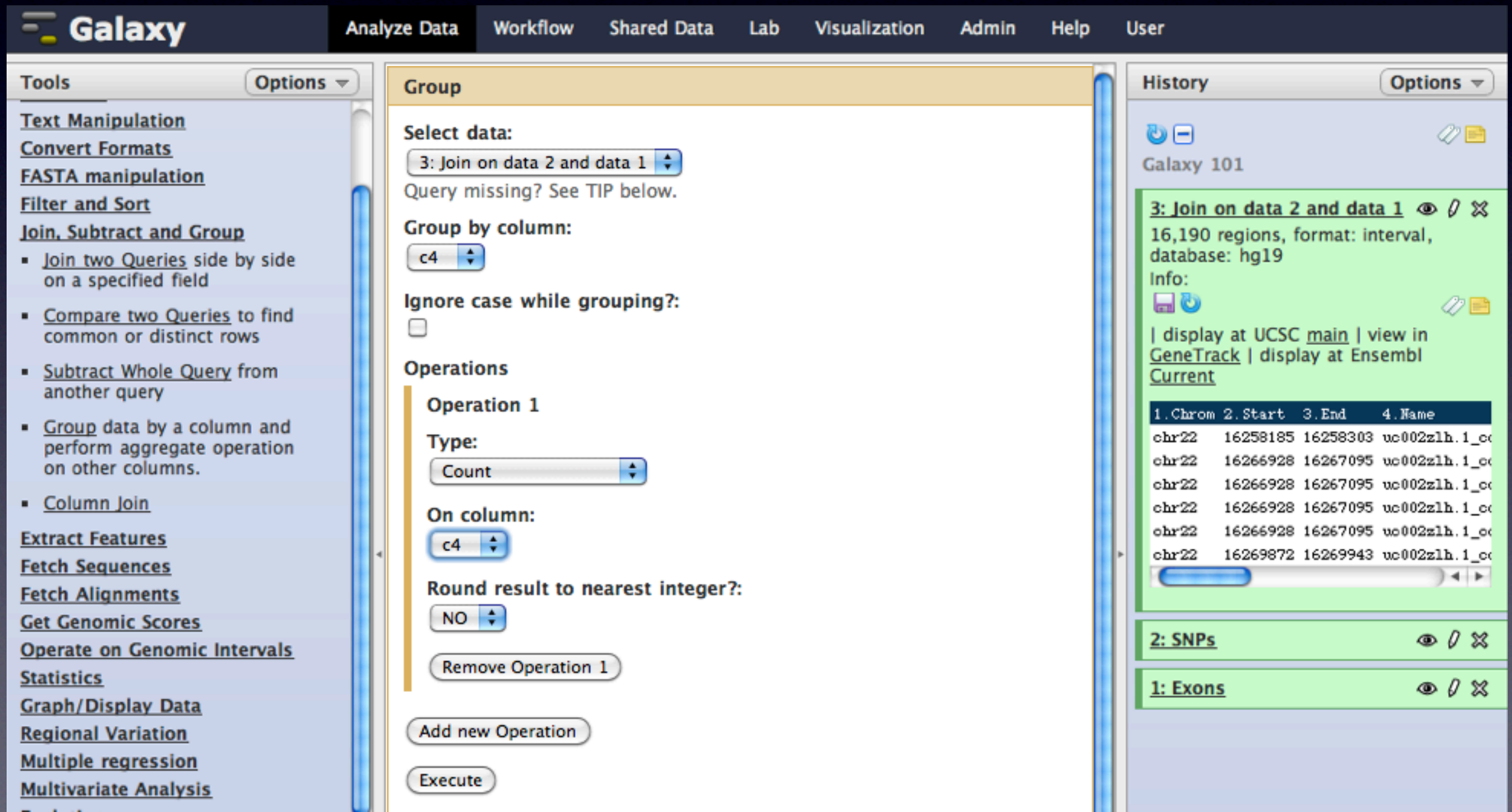
Join Data


Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
Tools
Options
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Intersect the intervals of two queries
Subtract the intervals of two queries
Merge the overlapping intervals of a query
Concatenate two queries into one query
Base Coverage of all intervals
Coverage of a set of intervals on second set of intervals
Complement intervals of a query
Cluster the intervals of a query
Join the intervals of two queries side-by-side
Get flanks returns flanking
Join
Join:
1: Exons
First query
with:
2: SNPs
Second query
with min overlap:
1
(bp)
Return:
Only records that are joined (INNER JOIN)
Execute
TIP: If your query does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.
Screencasts!
See Galaxy Interval Operation Screencasts (right click to open this link in another window).
Syntax
Where overlap specifies the minimum overlap between intervals that allows them to be joined.
Return only records that are joined returns only the records of the
History
Options
Galaxy 101
3: Join on data 2 and data 1
2: SNPs
1: Exons

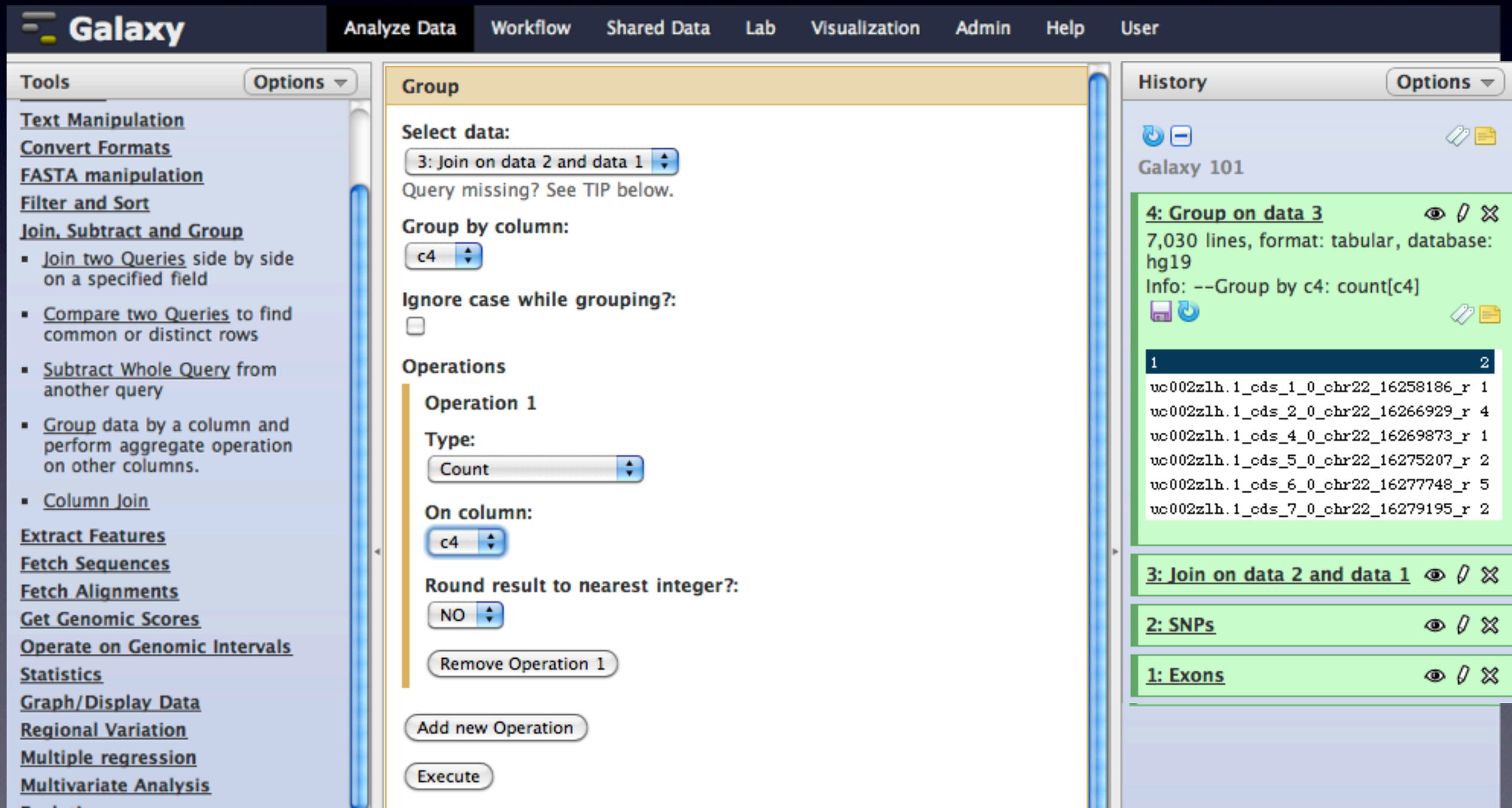
Manipulations



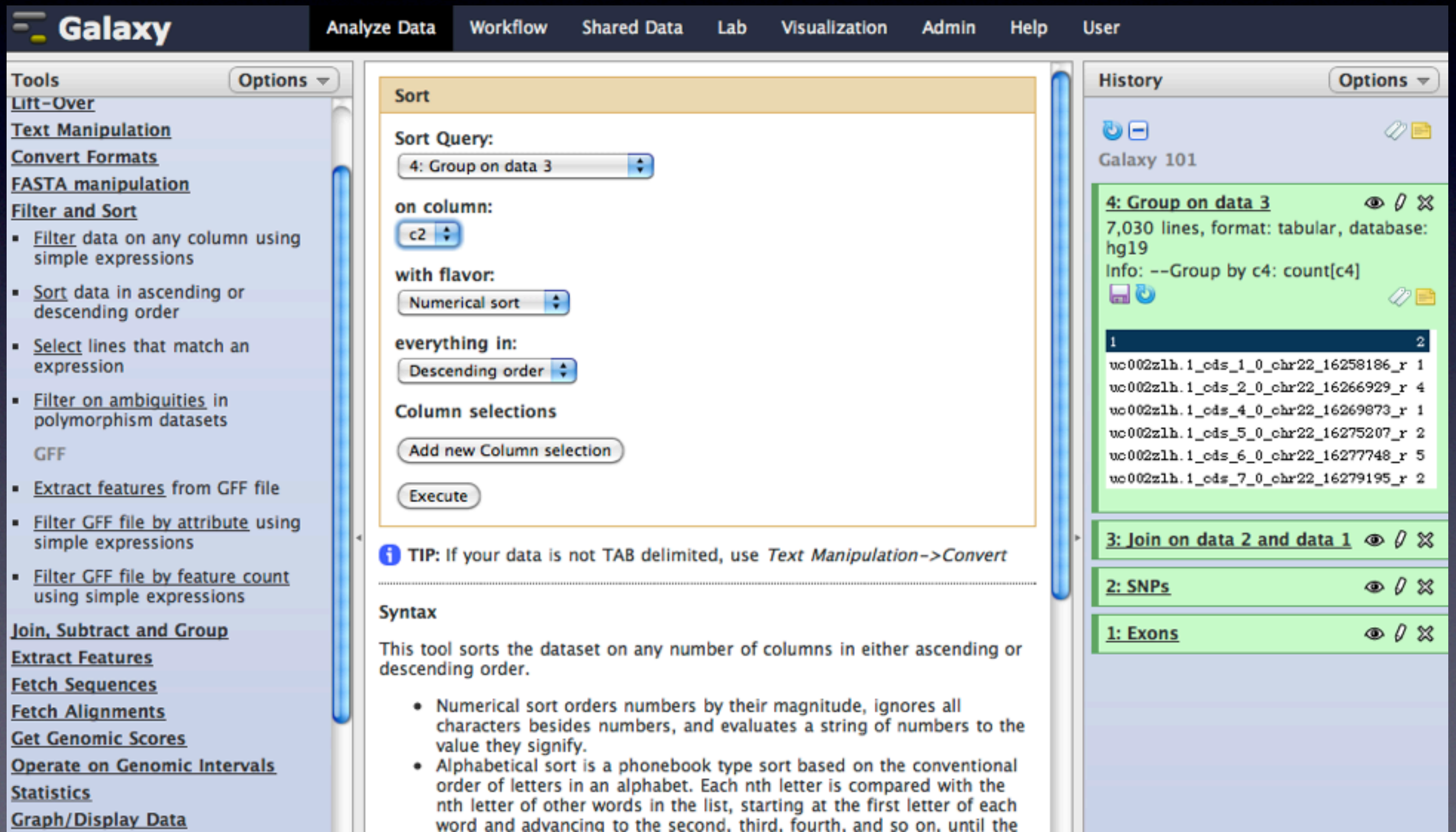
Manipulations



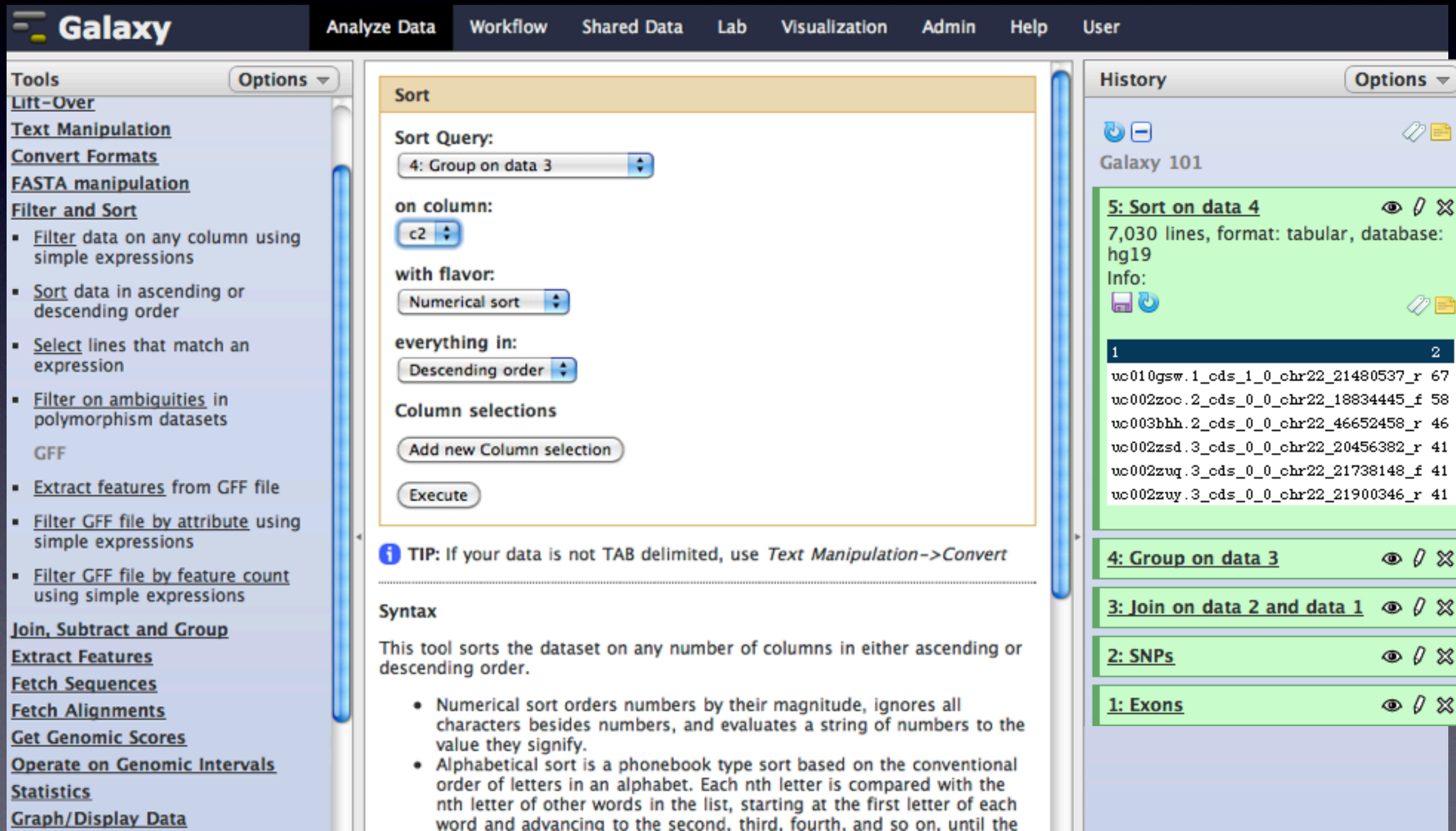
Manipulations



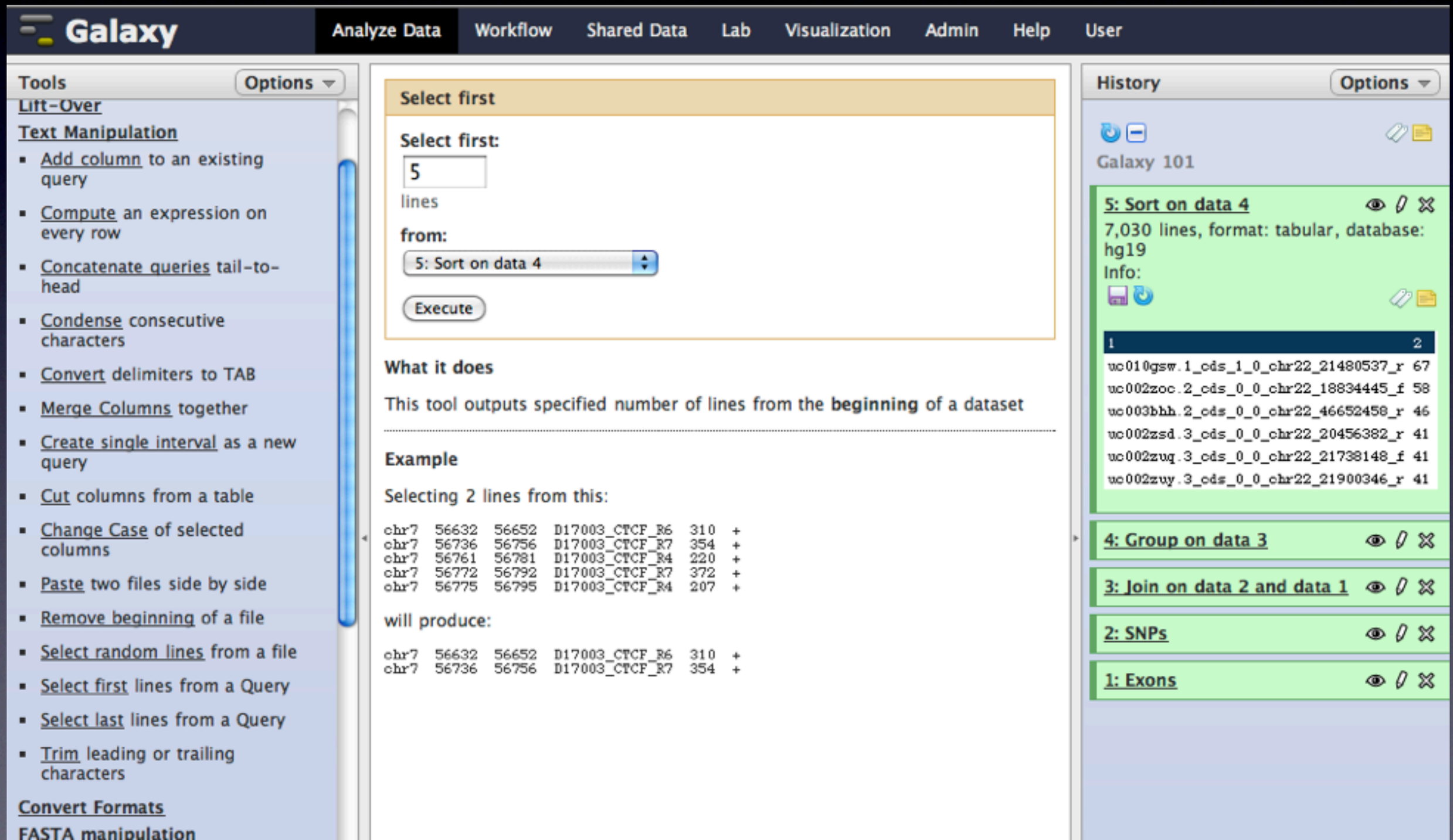
Manipulations



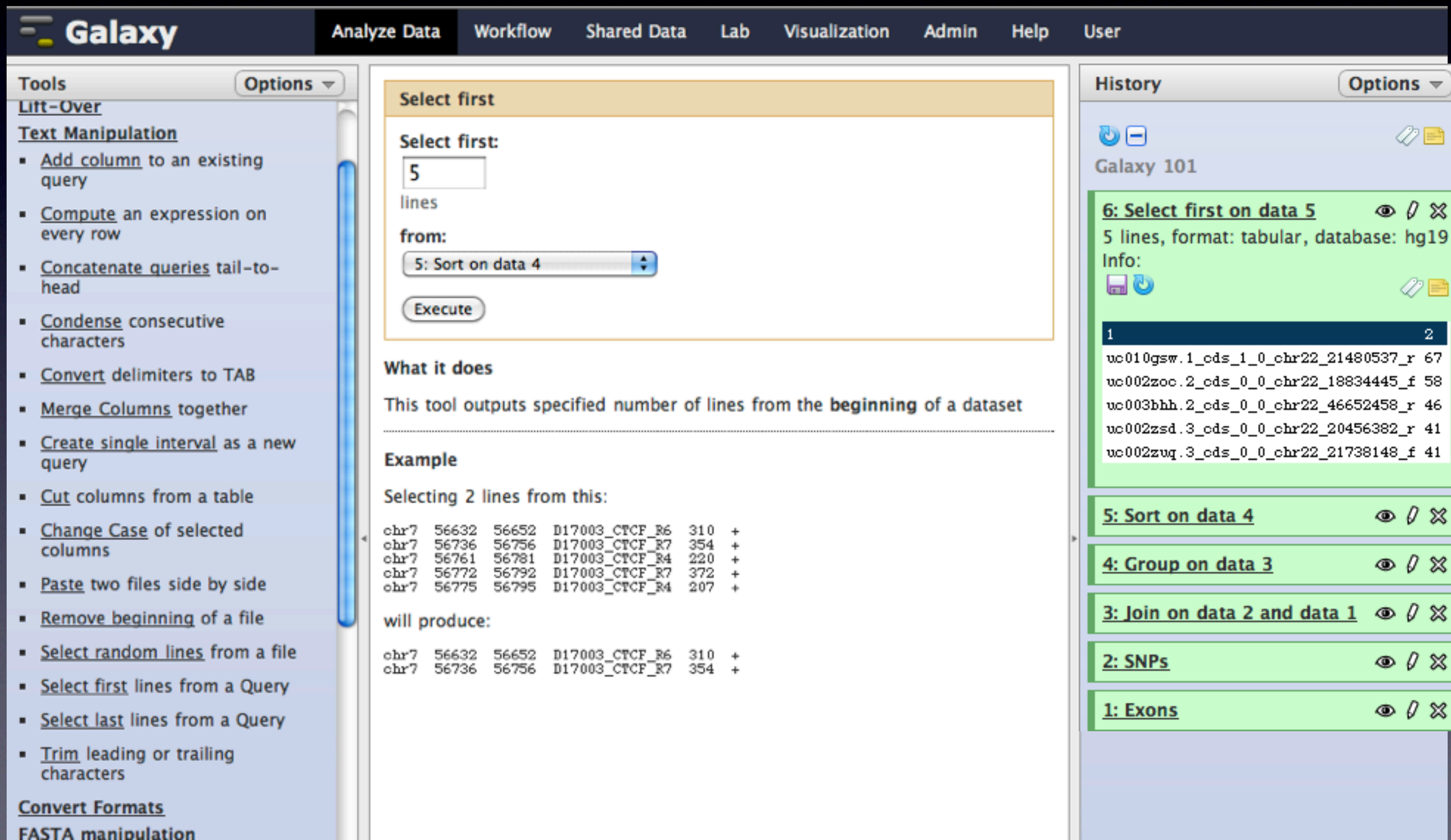
Manipulations



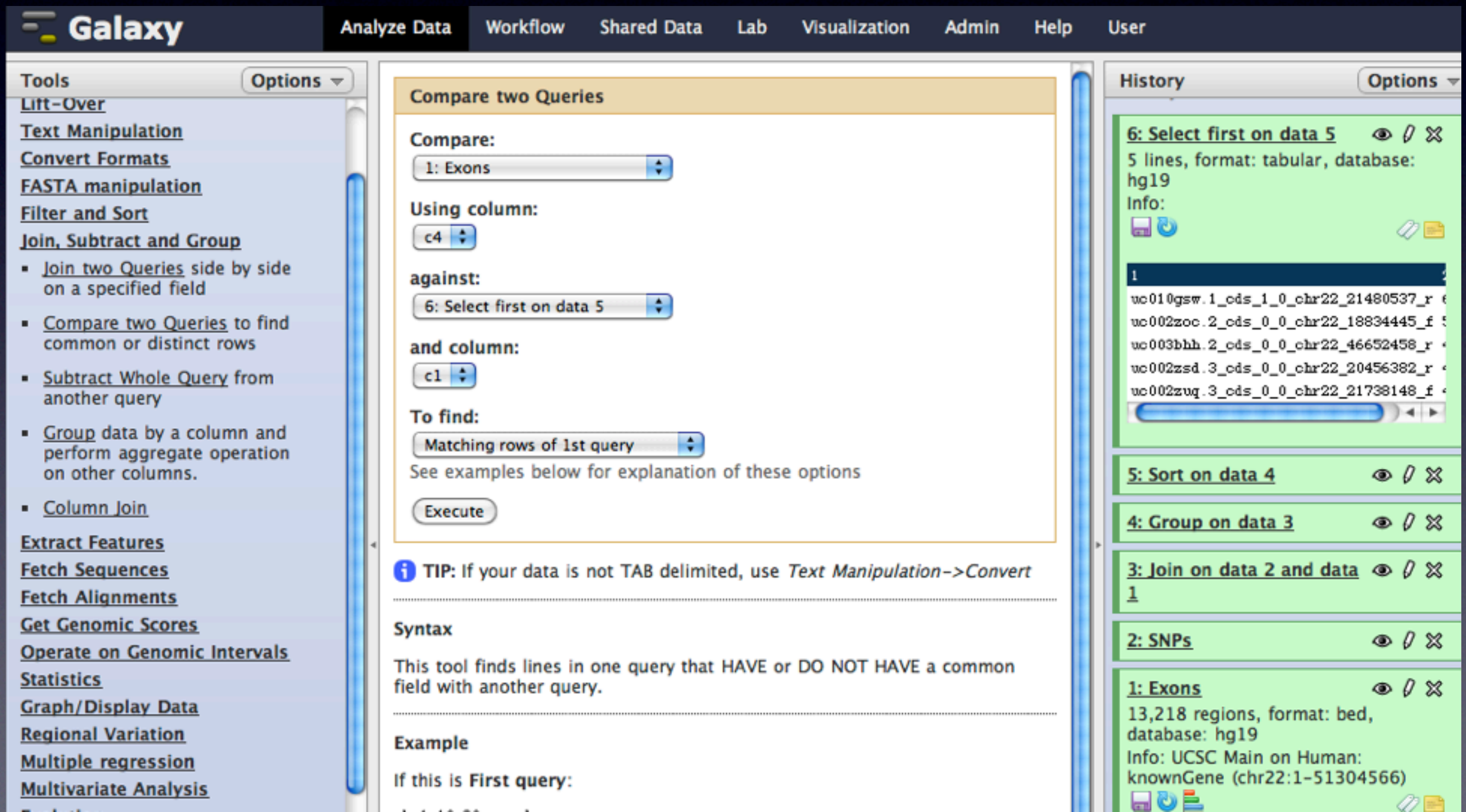
Manipulations



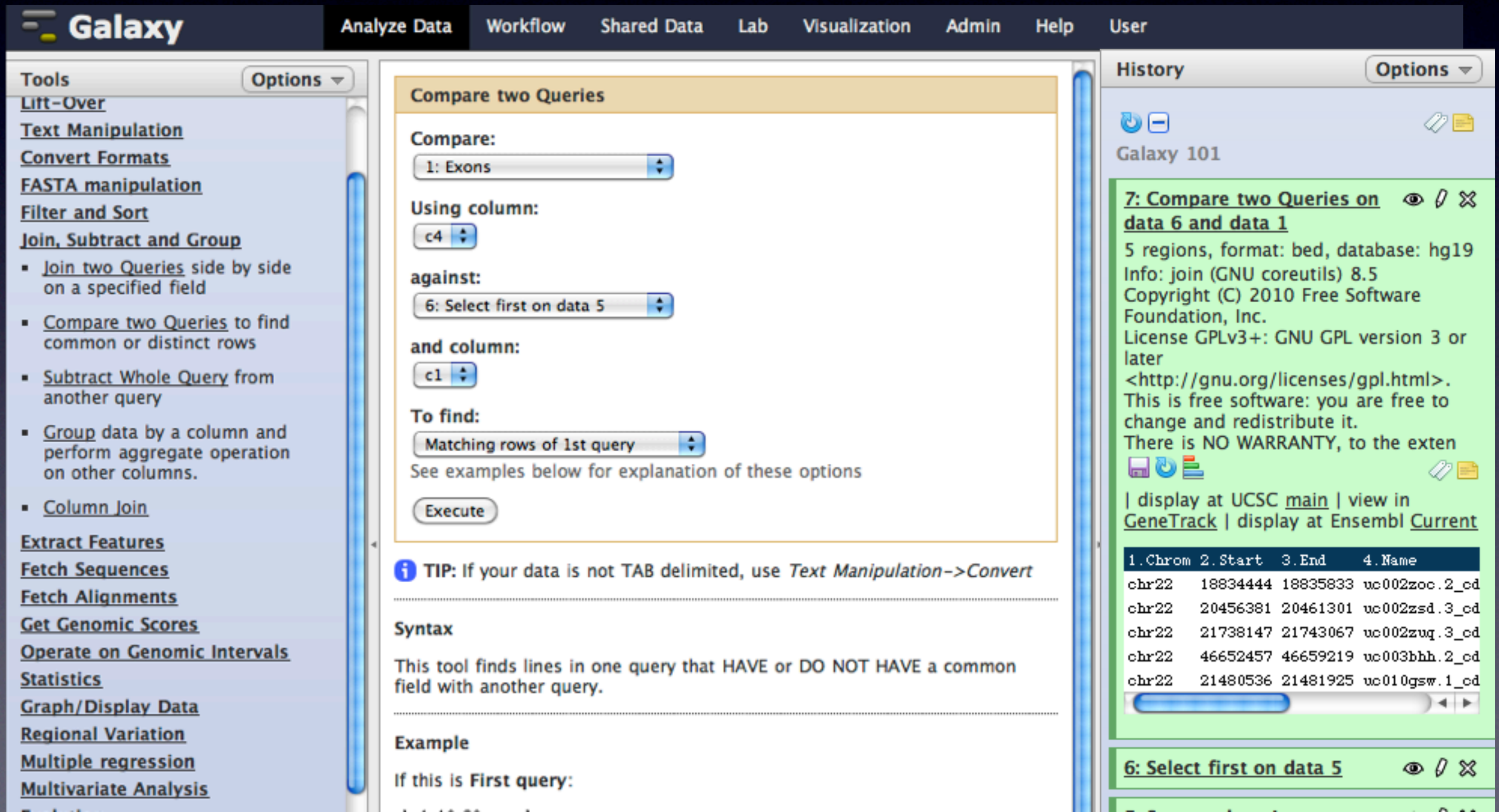
Manipulations



Display


Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
Tools
Options
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Join two Queries side by side on a specified field
Compare two Queries to find common or distinct rows
Subtract Whole Query from another query
Group data by a column and perform aggregate operation on other columns.
Column Join
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Compare two Queries
Compare:
1: Exons
Using column:
c4
against:
6: Select first on data 5
and column:
c1
To find:
Matching rows of 1st query
See examples below for explanation of these options
Execute
TIP: If your data is not TAB delimited, use Text Manipulation->Convert
Syntax
This tool finds lines in one query that HAVE or DO NOT HAVE a common field with another query.
Example
If this is First query:
History
Options
6: Select first on data 5
5 lines, format: tabular, database: hg19
Info:
uc010gsw.1_cds_1_0_chr22_21480537_r
uc002zoc.2_cds_0_0_chr22_18834445_f
uc003bhh.2_cds_0_0_chr22_46652458_r
uc002zsd.3_cds_0_0_chr22_20456382_r
uc002zuq.3_cds_0_0_chr22_21738148_f
5: Sort on data 4
4: Group on data 3
3: Join on data 2 and data 1
2: SNPs
1: Exons
13,218 regions, format: bed, database: hg19
Info: UCSC Main on Human: knownGene (chr22:1-51304566)

Display


Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
Tools
Options
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Join two Queries side by side on a specified field
Compare two Queries to find common or distinct rows
Subtract Whole Query from another query
Group data by a column and perform aggregate operation on other columns.
Column Join
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Compare two Queries
Compare:
1: Exons
Using column:
c4
against:
6: Select first on data 5
and column:
c1
To find:
Matching rows of 1st query
See examples below for explanation of these options
Execute
TIP: If your data is not TAB delimited, use Text Manipulation->Convert
Syntax
This tool finds lines in one query that HAVE or DO NOT HAVE a common field with another query.
Example
If this is First query:
History
Options
Galaxy 101
7: Compare two Queries on data 6 and data 1
5 regions, format: bed, database: hg19
Info: join (GNU coreutils) 8.5
Copyright (C) 2010 Free Software Foundation, Inc.
License GPLv3+: GNU GPL version 3 or later
<http://gnu.org/licenses/gpl.html>.
This is free software: you are free to change and redistribute it.
There is NO WARRANTY, to the exten
| display at UCSC main | view in GeneTrack | display at Ensembl Current
1.Chrom 2.Start 3.End 4.Name
chr22 18834444 18835833 uc002zoc.2_cd
chr22 20456381 20461301 uc002zsd.3_cd
chr22 21738147 21743067 uc002zuq.3_cd
chr22 46652457 46659219 uc003bhh.2_cd
chr22 21480536 21481925 uc010gsw.1_cd
6: Select first on data 5

Galaxy 101
Live Demonstration

- Show you how to perform an experiment
- Obtain data from external sources
- Perform data manipulation and display data
- Understand Galaxy's History system
- Create, modify, apply workflows to data

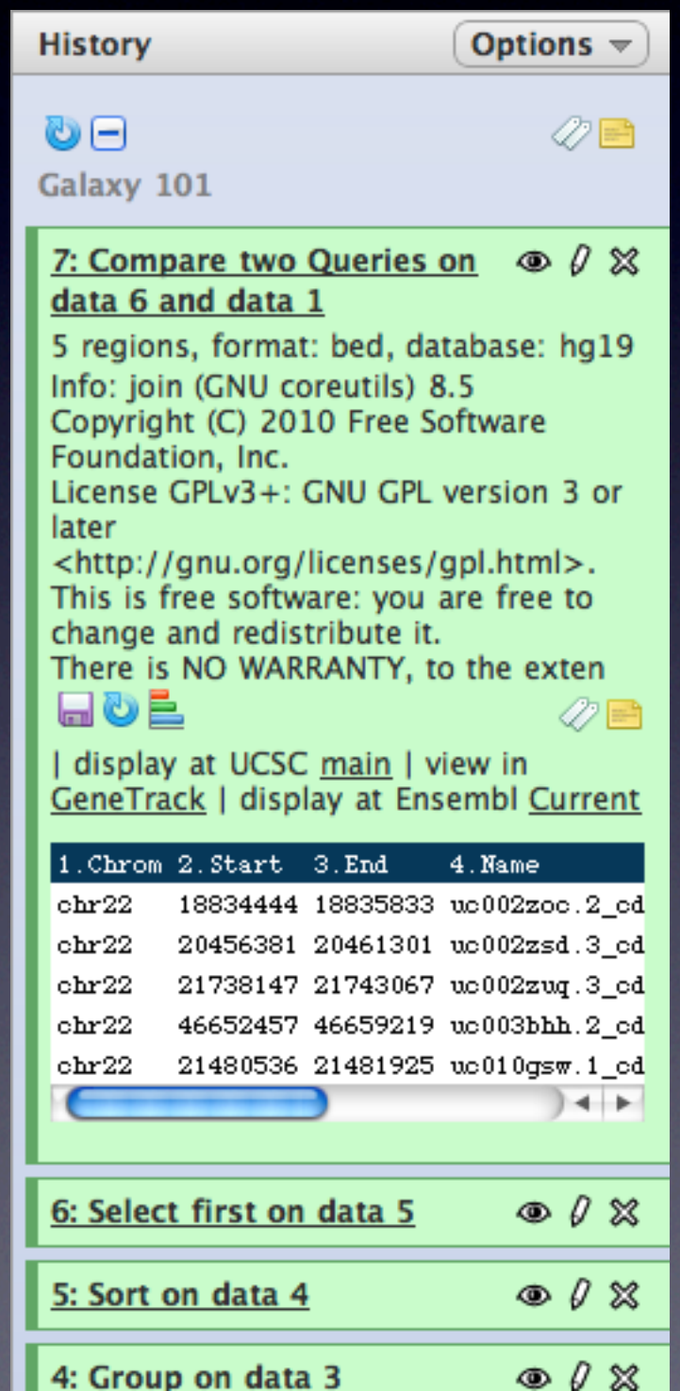
based on A. Nekrutenko et al. (2010) Galaxy 101 Tutorial
https://main.g2.bx.psu.edu/galaxy101

Galaxy 101
Live Demonstration

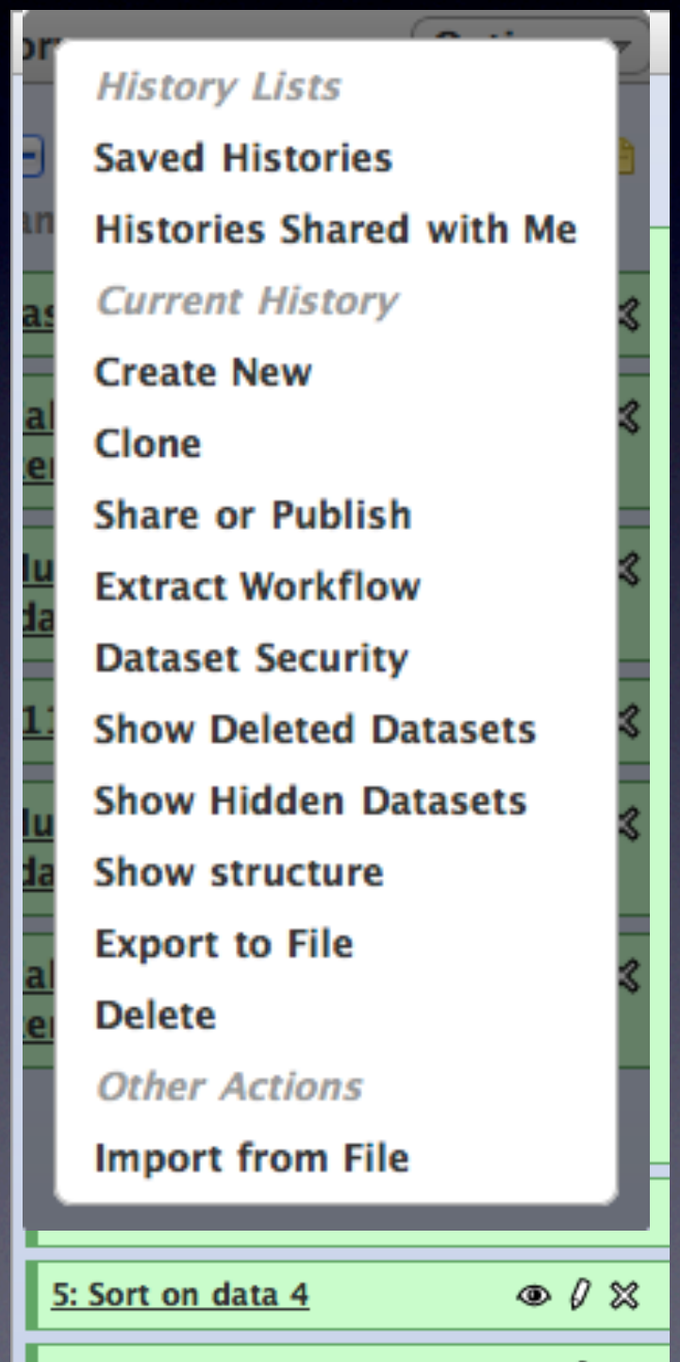
- Show you how to perform an experiment
- Obtain data from external sources
- Perform data manipulation and display data
- Understand Galaxy's History system
- Create, modify, apply workflows to data

based on A. Nekrutenko et al. (2010) Galaxy 101 Tutorial
https://main.g2.bx.psu.edu/galaxy101

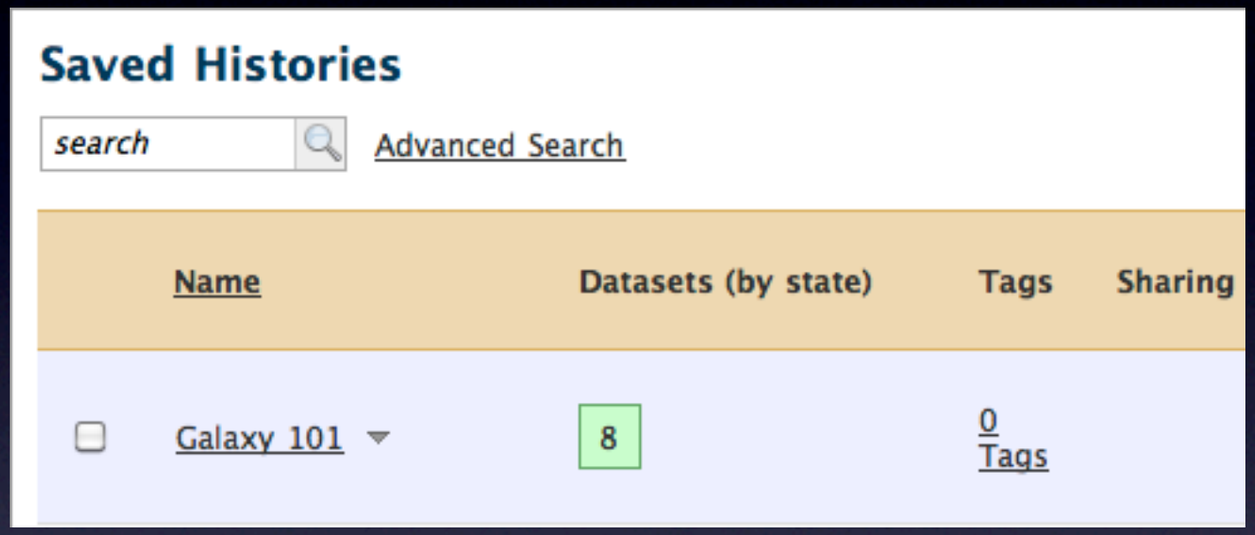
History


History
Options
Galaxy 101
7: Compare two Queries on data 6 and data 1
5 regions, format: bed, database: hg19
Info: join (GNU coreutils) 8.5
Copyright (C) 2010 Free Software Foundation, Inc.
License GPLv3+: GNU GPL version 3 or later
<http://gnu.org/licenses/gpl.html>.
This is free software: you are free to change and redistribute it.
There is NO WARRANTY, to the exten
| display at UCSC main | view in GeneTrack | display at Ensembl Current
1.Chrom 2.Start 3.End 4.Name
chr22 18834444 18835833 uc002zoc.2_cd
chr22 20456381 20461301 uc002zsd.3_cd
chr22 21738147 21743067 uc002zuq.3_cd
chr22 46652457 46659219 uc003bhh.2_cd
chr22 21480536 21481925 uc010gsw.1_cd
6: Select first on data 5
5: Sort on data 4
4: Group on data 3

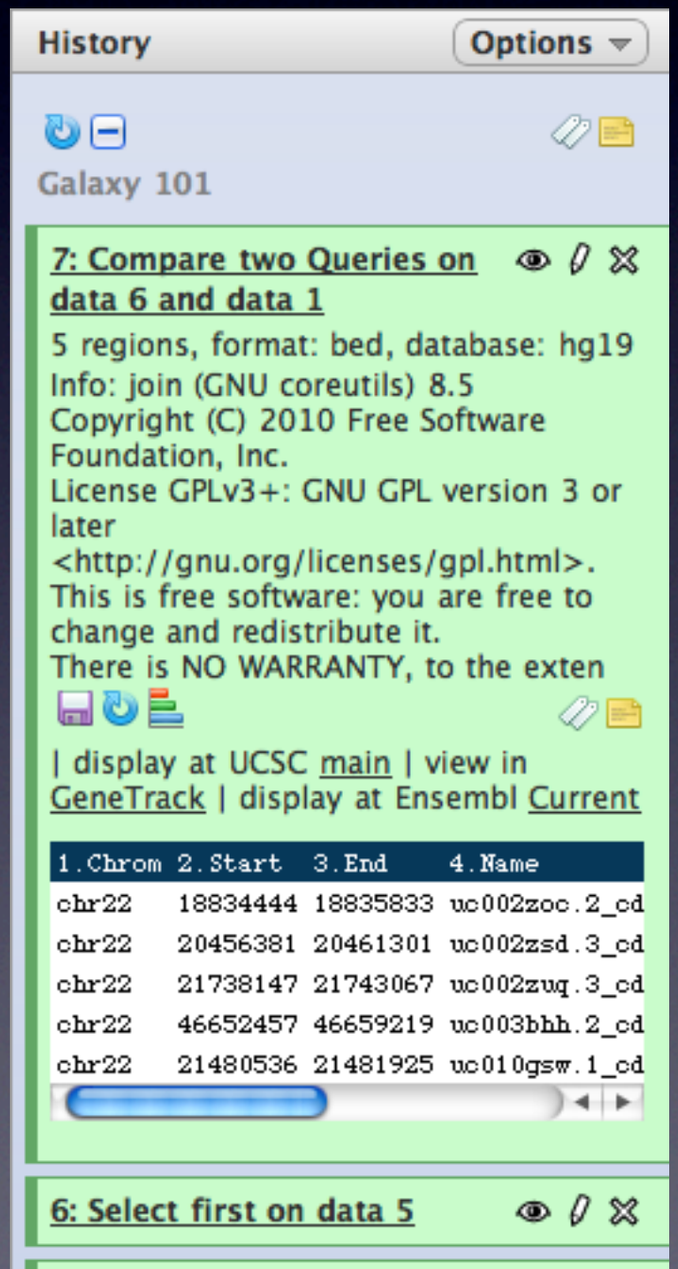
History


History Lists
Saved Histories
Histories Shared with Me
Current History
Create New
Clone
Share or Publish
Extract Workflow
Dataset Security
Show Deleted Datasets
Show Hidden Datasets
Show structure
Export to File
Delete
Other Actions
Import from File
5: Sort on data 4
4: Group on data 3

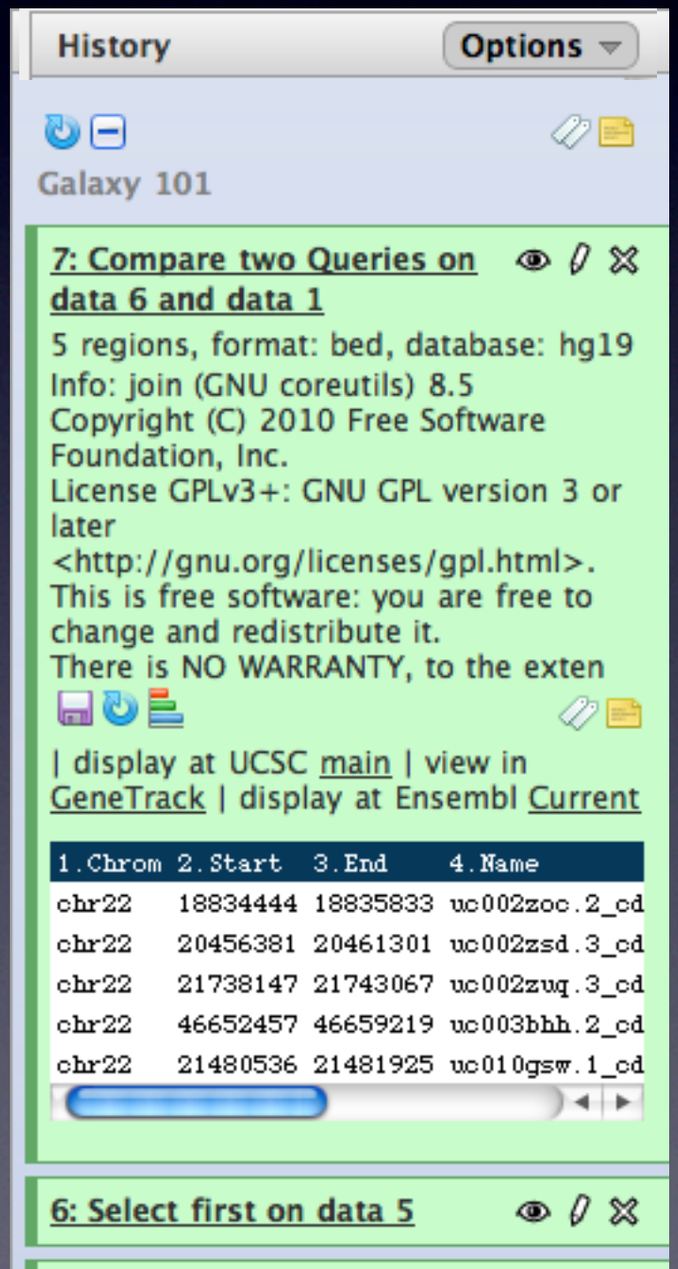
History


Saved Histories
search
Advanced Search
Name
Datasets (by state)
Tags
Sharing
Galaxy 101
8
0
Tags

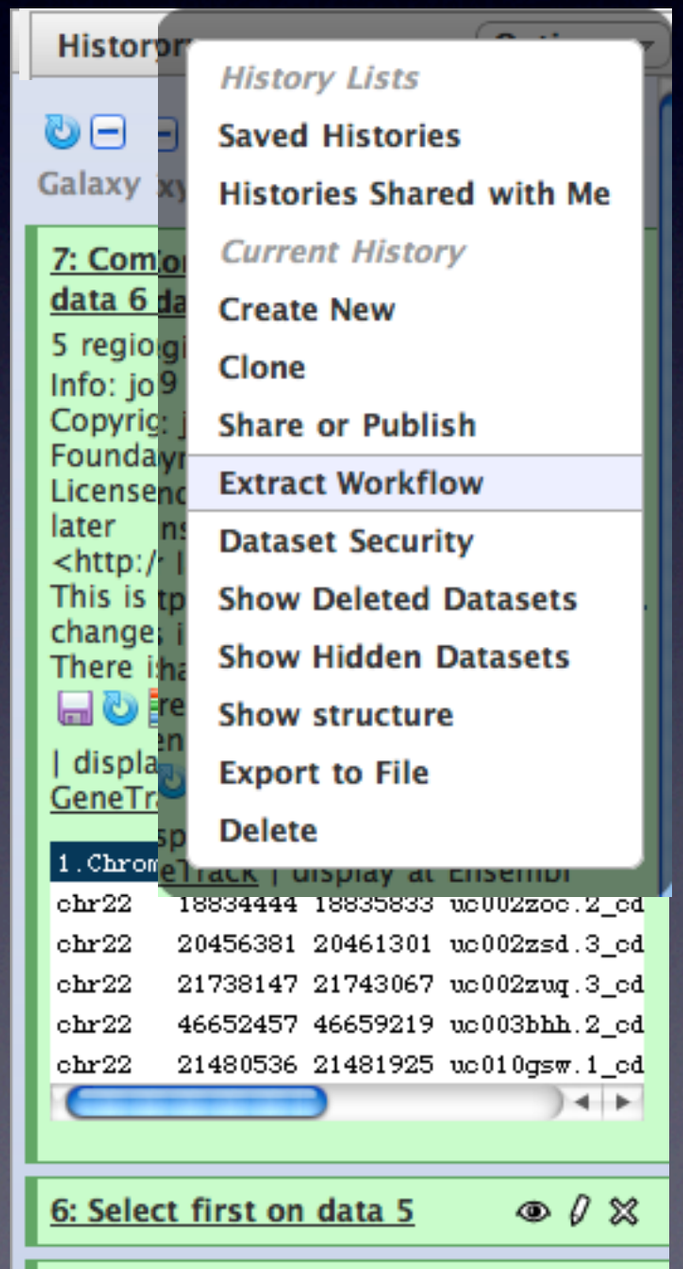
Extracting a Workflow



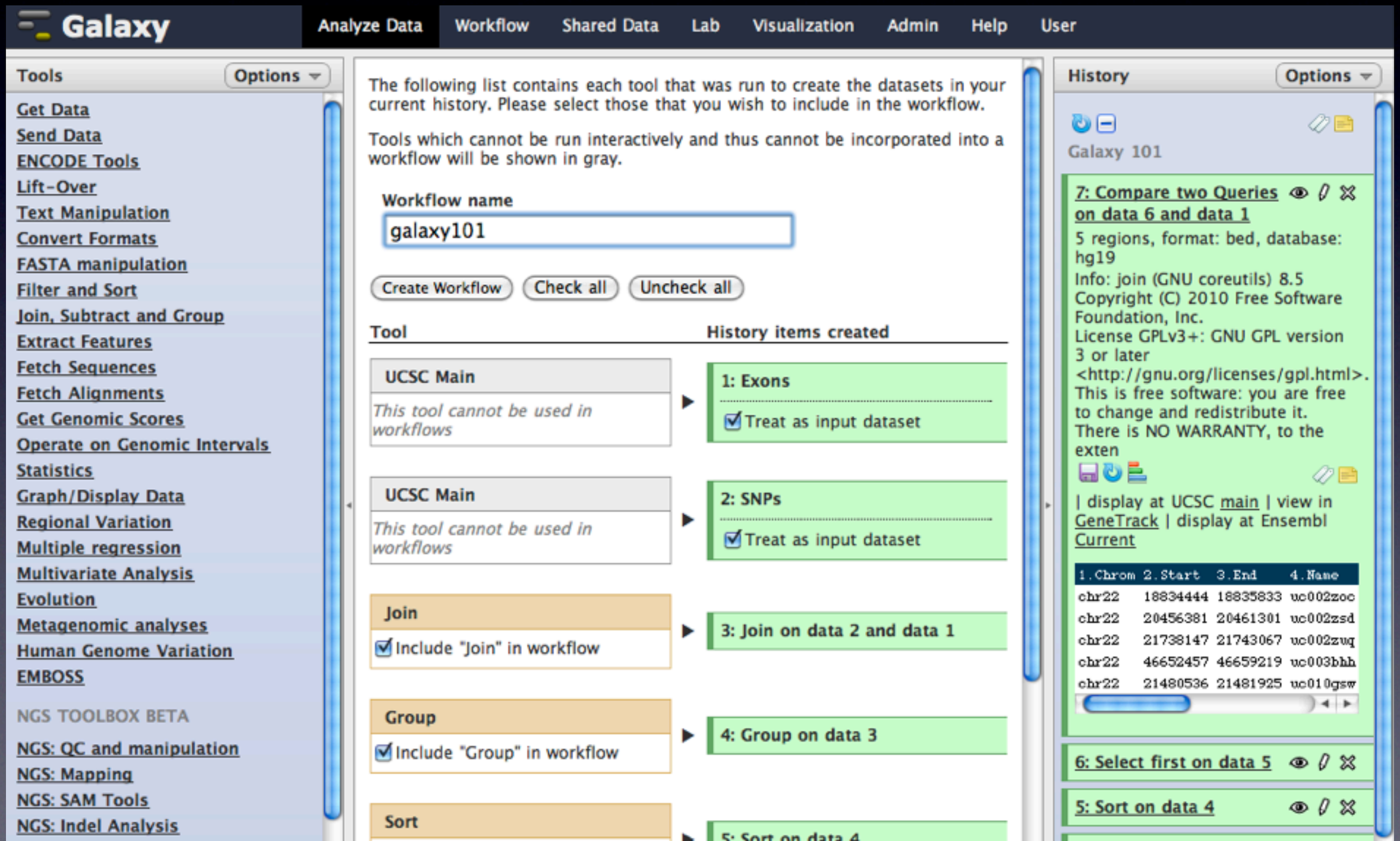
Extracting a Workflow



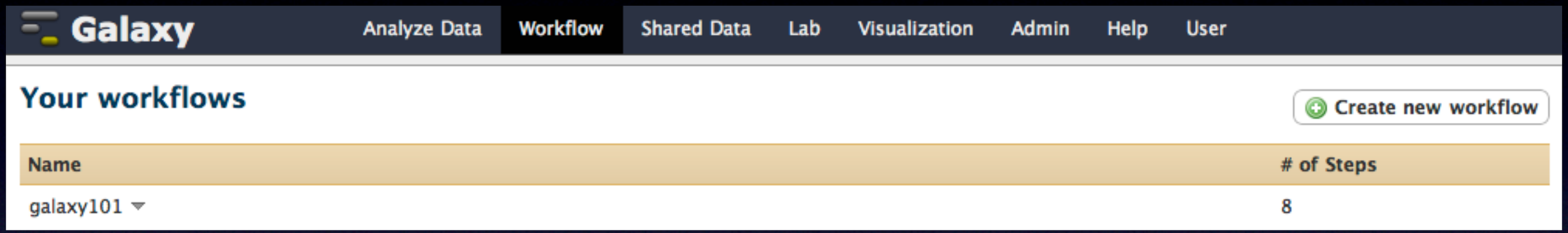
Extracting a Workflow



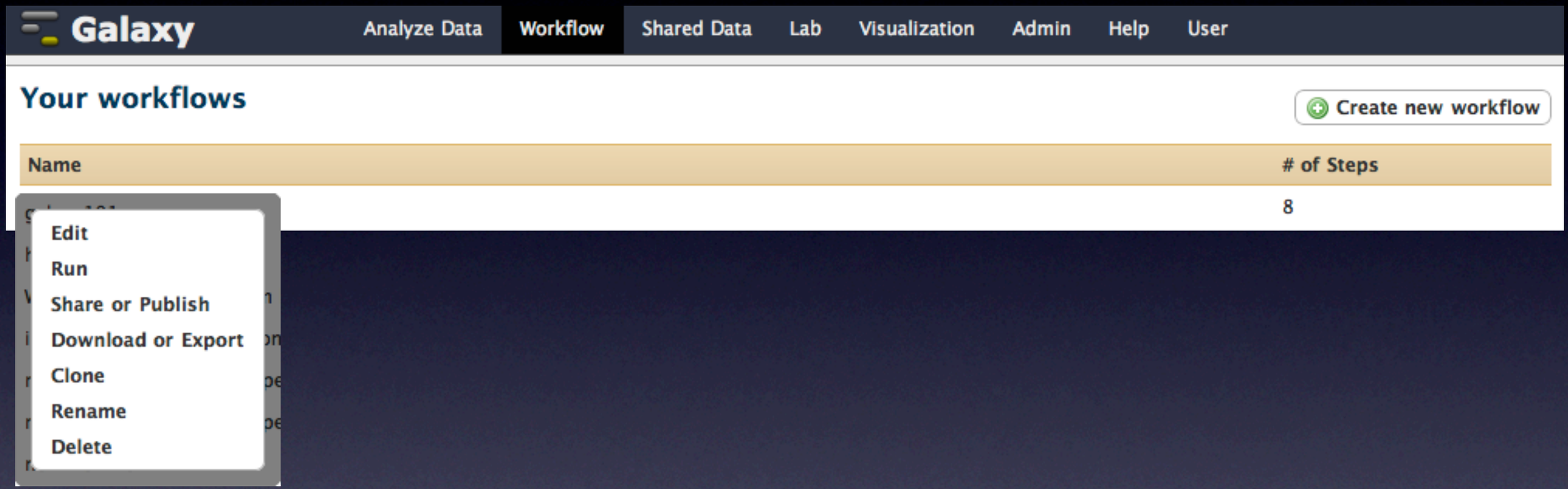
Extracting a Workflow



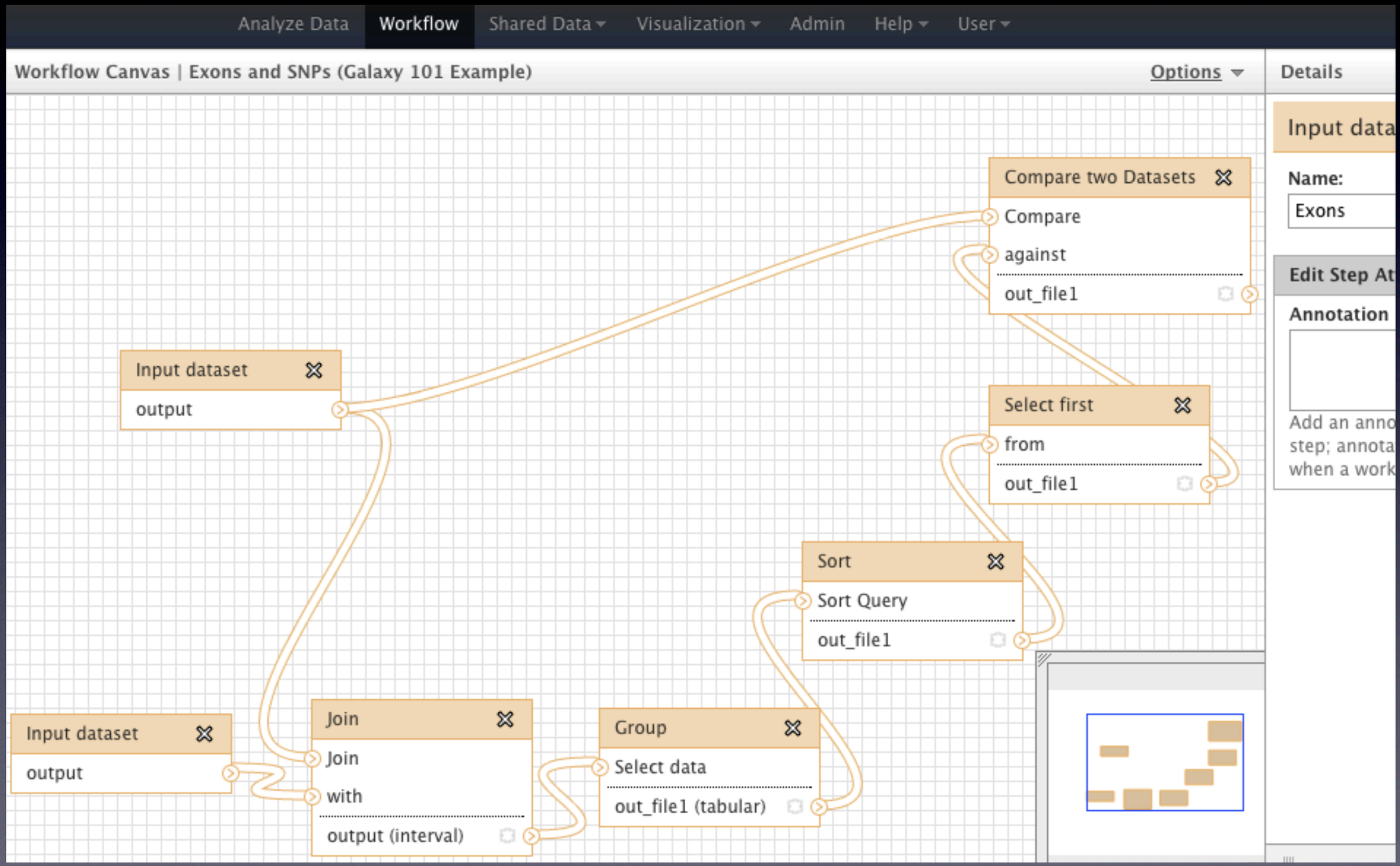
Extracting a Workflow



Extracting a Workflow



Editing a Workflow

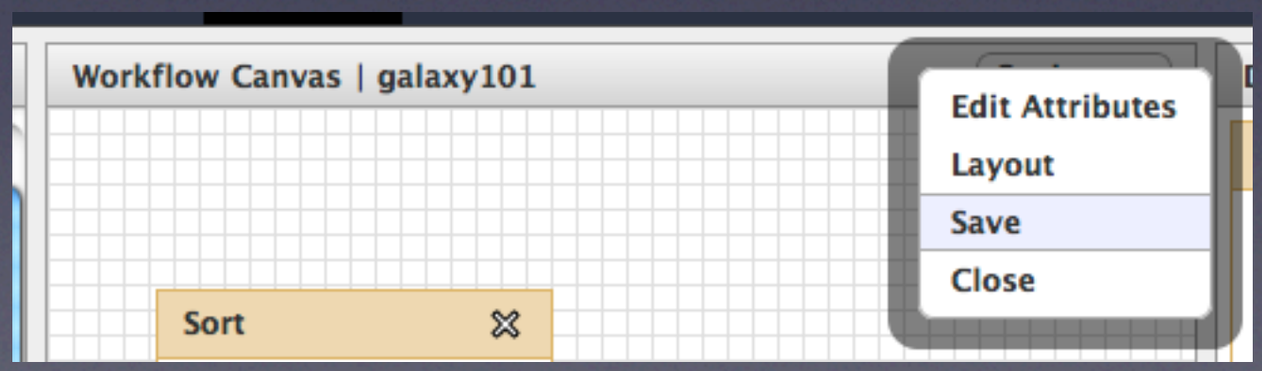


You can also …

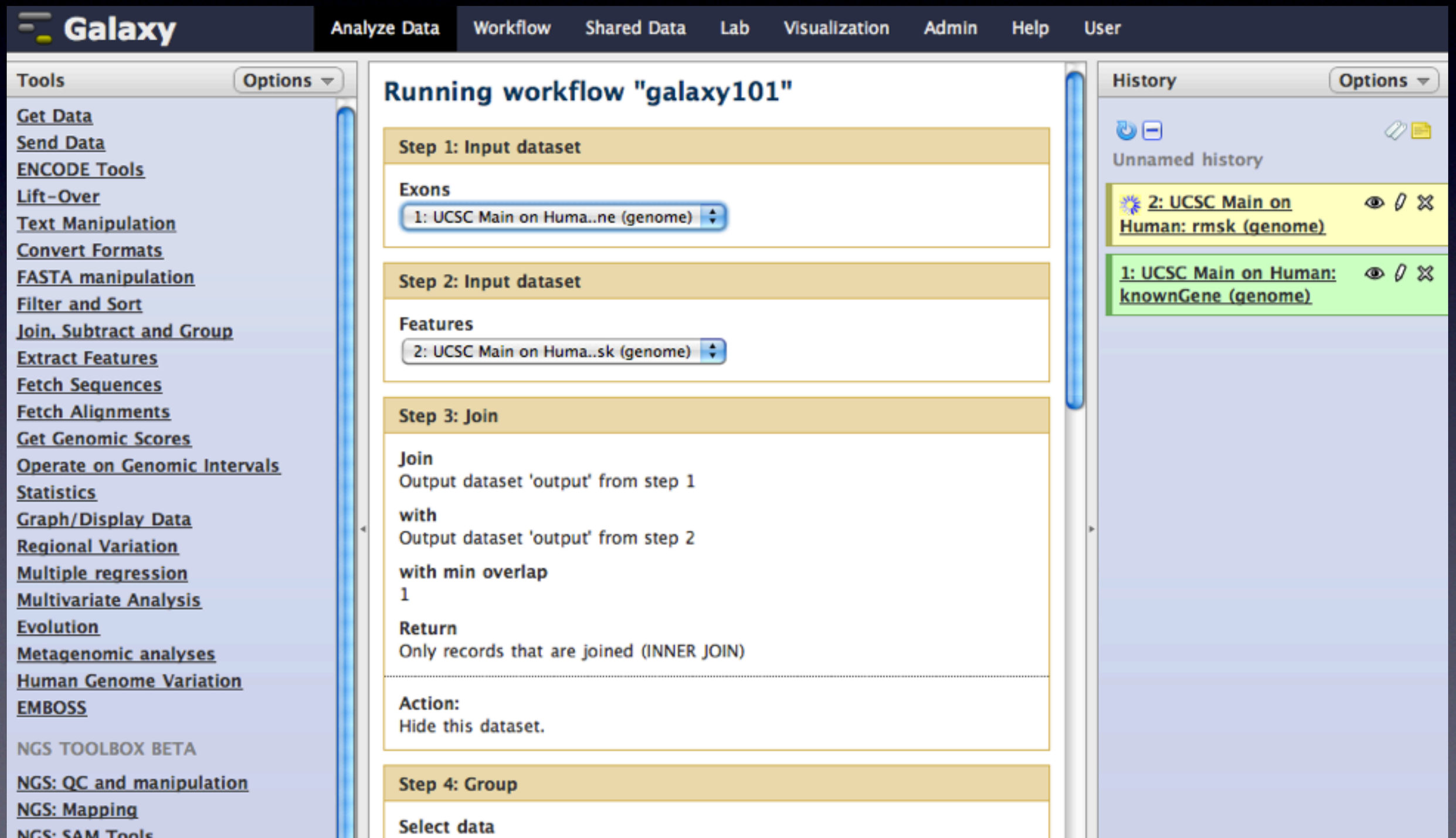
- Rename inputs & outputs
- Hide intermediate steps
- Set default and user-editable parameters
- Save your work!

You can also …

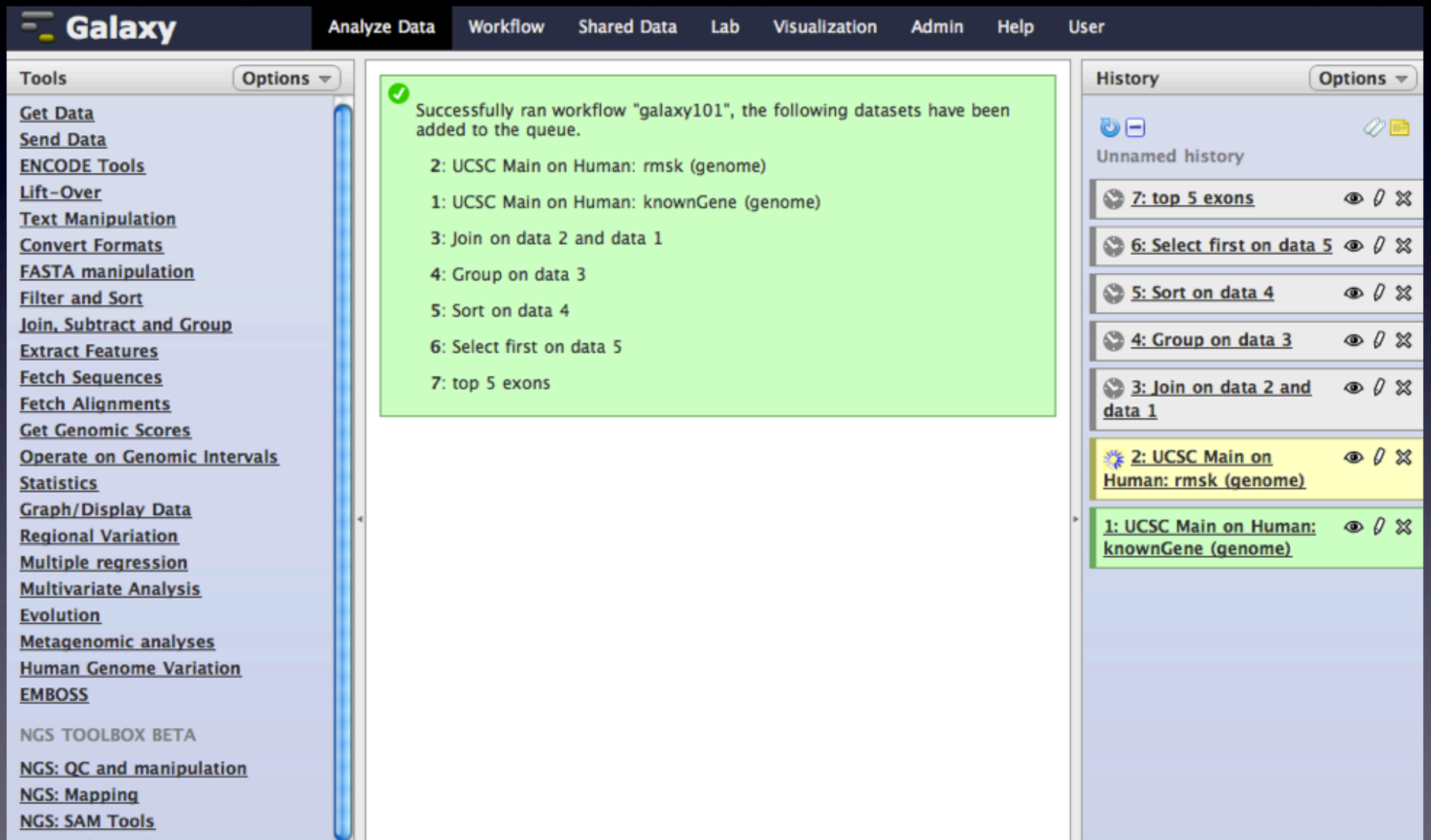
- Rename inputs & outputs
- Hide intermediate steps
- Set default and user-editable parameters
- Save your work!



Run Workflow on New Data



Wait for it ...


Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
Tools
Options
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
Human Genome Variation
EMBOSS
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
Successfully ran workflow "galaxy101", the following datasets have been added to the queue.
2: UCSC Main on Human: rmsk (genome)
1: UCSC Main on Human: knownGene (genome)
3: Join on data 2 and data 1
4: Group on data 3
5: Sort on data 4
6: Select first on data 5
7: top 5 exons
History
Options
Unnamed history
7: top 5 exons
6: Select first on data 5
5: Sort on data 4
4: Group on data 3
3: Join on data 2 and data 1
2: UCSC Main on Human: rmsk (genome)
1: UCSC Main on Human: knownGene (genome)

Galaxy 101
Live Demonstration

- Show you how to perform an experiment
- Obtain data from external sources
- Perform data manipulation and display data
- Understand Galaxy's History system
- Create, modify, apply workflows to data

based on A. Nekrutenko et al. (2010) Galaxy 101 Tutorial
https://main.g2.bx.psu.edu/galaxy101

Galaxy 101
Live Demonstration

- Show you how to perform an experiment
- Obtain data from external sources
- Perform data manipulation and display data
- Understand Galaxy's History system
- Create, modify, apply workflows to data

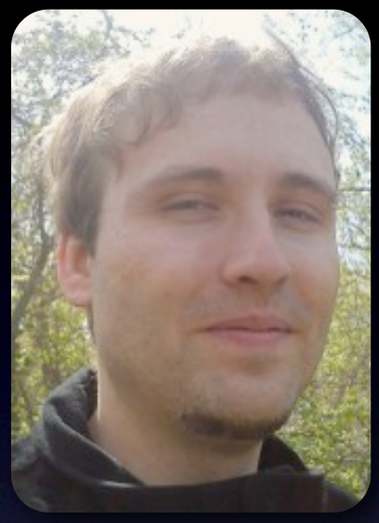
based on A. Nekrutenko et al. (2010) Galaxy 101 Tutorial
https://main.g2.bx.psu.edu/galaxy101

Conclusions

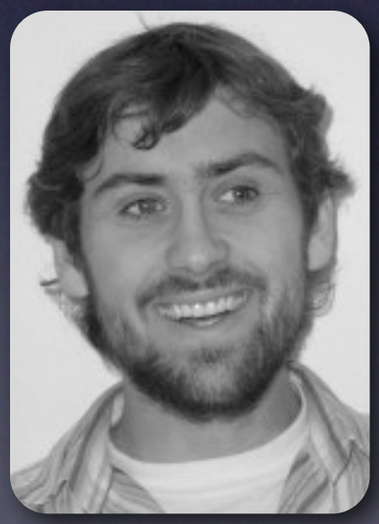
- Share workflows to standardize analyses
- NGS: Need for distributed computing
- Part II: oqtans tomorrow

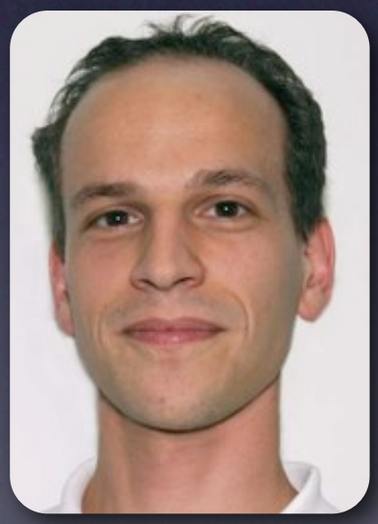
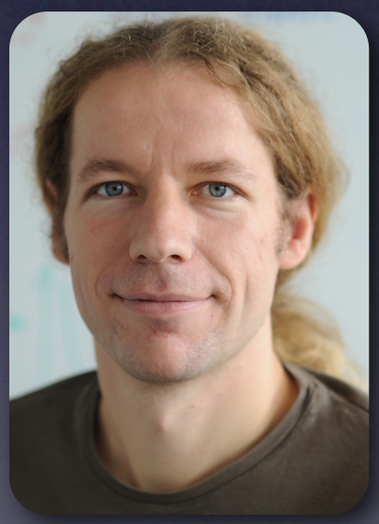
http://oqtans.org



Jonas Behr, Regina Bohnert, Philipp Drewe, Nico Görnitz



André Kahles, Pramod Mudrakarta, Vipin T. Sreedharan, Georg Zeller, Gunnar Rätsch