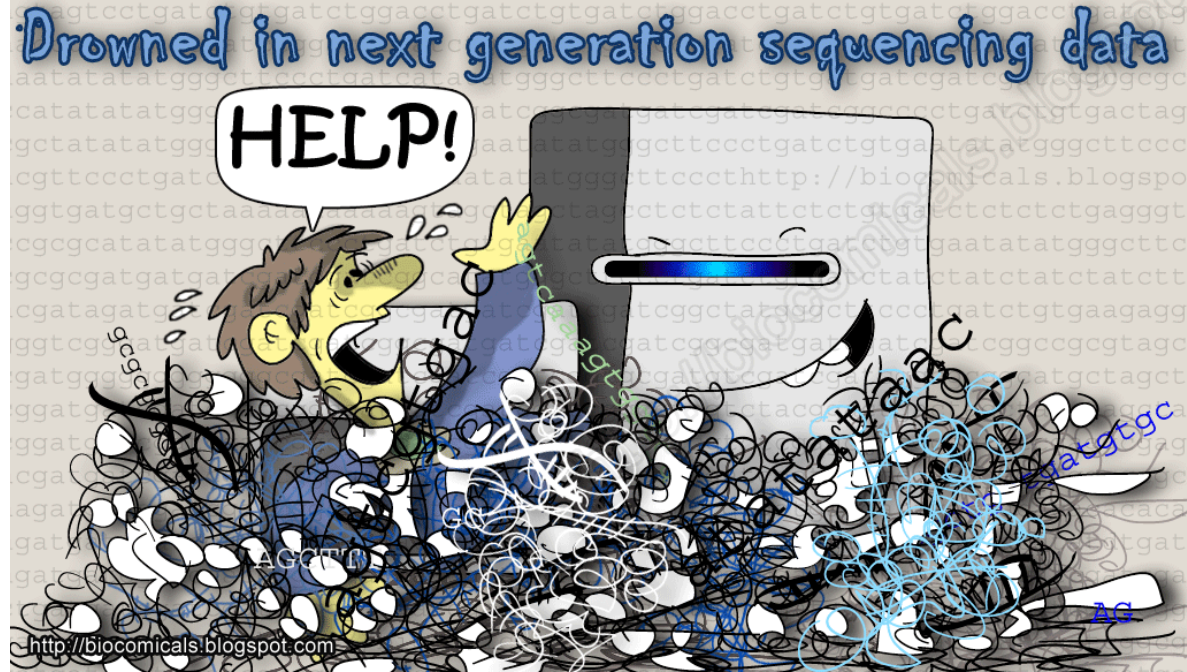


Popular bioinformatics tools in Galaxy: I

Yanbin Yin
Spring 2013

As creating and obtaining data has become easier, the key decision faced by many researchers is a practical one: where and how should an analysis be performed?



1. No need of programming experience.
2. Integrates many bioinformatics tools within one interface.
3. Keeps track of all the steps performed in an analysis. Even if you delete the datasets, the history keeps the tools used.

Galaxy (<http://galaxyproject.org>) is a **software system** that provides genomics data analysis support through a framework that gives experimentalists **simple interfaces** to powerful **computational tools**, while automatically managing the computational details.

Penn State U:
Anton Nekrutenko



Emory U:
James Taylor



Bioinformatics
Tools and analyses:

Accessibility
Reproducibility
Transparency



Galaxy is a metasever that allows users to:

- retrieve information from multiple remote sources
- store, combine, and refine the information at a central site
- perform mathematical operations
- analyze the results using sophisticated tools

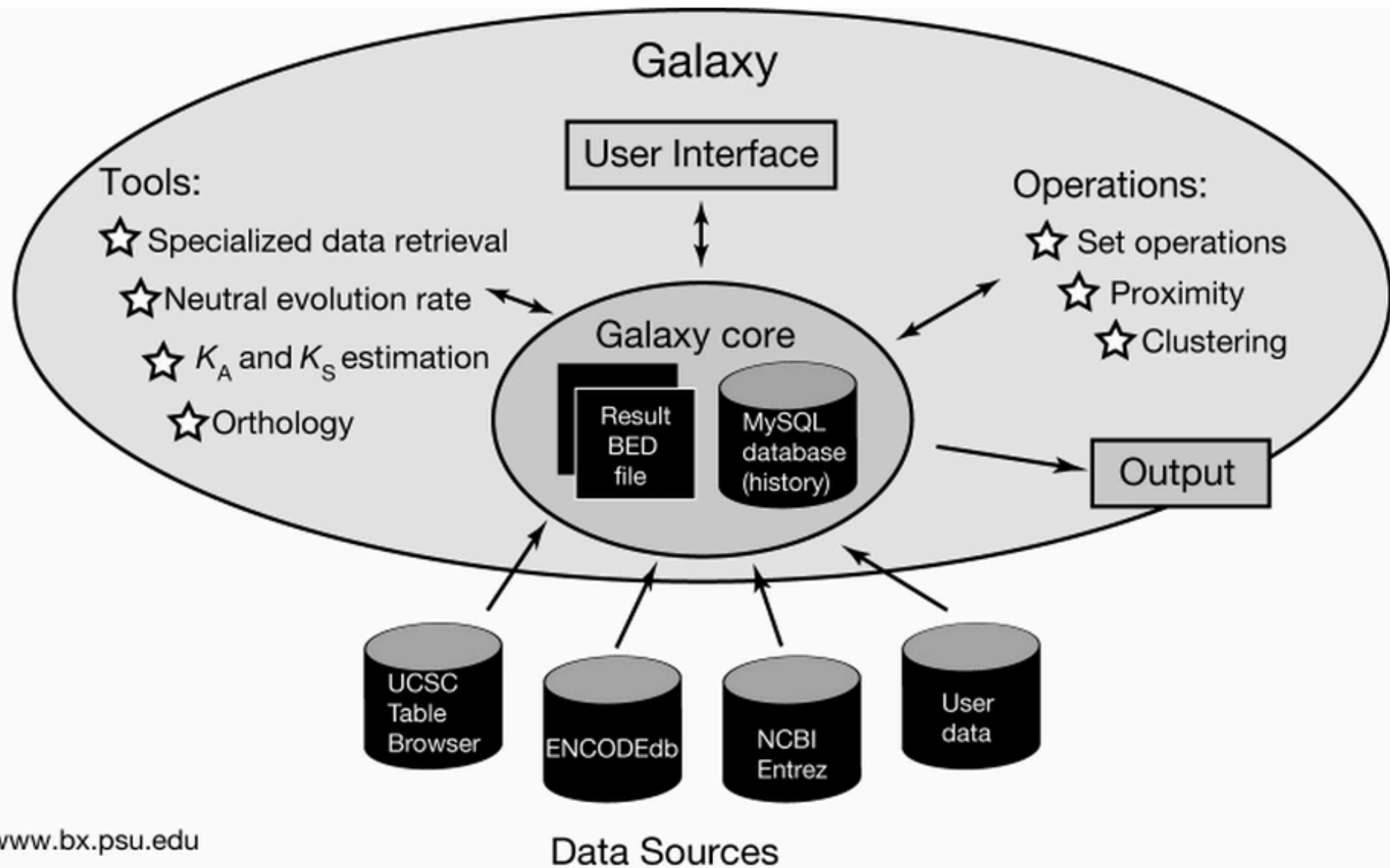


Galaxy is designed for the integration of:

- genomic sequences
- alignments of those sequences
- functional annotations



Galaxy combines the power of existing databases and visualization engines with seamless access to a wide variety of analytical tools.



Hands on!

Data intensive biology *for everyone.*

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.

Use Galaxy



Use the free public server

Get Galaxy



Install locally or in the cloud

Learn Galaxy



Screencasts, Galaxy 101, ...

Get Involved

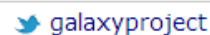


Mailing lists, Tool Shed, wiki

[Search all resources](#)

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.

MCC. U-...-A-B-A-A-...



8

Data intensive biology *for everyone.*

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.

Use Galaxy



Use the free public server

Get Galaxy



Install locally or in the cloud

Learn Galaxy



Screencasts, Galaxy 101, ...

Get Involved



Mailing lists, Tool Shed, wiki

[Search all resources](#)

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.

Get Galaxy: Galaxy Download and Installation

In addition to using the [public Galaxy server](#) (a.k.a. [Main](#)), you can also install your own instance of Galaxy (what this page is about), or create an [instance of Galaxy on the cloud](#). Another option is to use one of the ever-increasing number of [Public Galaxy Servers](#) hosted by other organizations.

See [Big Picture/Choices](#) for help on deciding which of these options may be best for your situation.

Reasons to Install Your Own Galaxy

You only need to download Galaxy if you plan to:

1. [Develop](#) it further
2. [Add](#) new tools
3. [Plug-in](#) new datasources, or
4. [Run](#) a local production server for your site because you have
 1. Sensitive data (e.g., clinical)
 2. Large datasets or processing requirements that are too big to be processed on [Main](#)

Installation Procedure

Contents

1. [Reasons to Install Your Own Galaxy](#)
2. [Installation Procedure](#)
 1. [Check your Python version](#)
 2. [Get the latest copy from the repository](#)
 3. [Start it up](#)
 4. [Join the Mailing List](#)
 5. [Keep your instance backed up](#)
 6. [Keep your code up to date](#)
3. [Advanced Configuration](#)
4. [Other Help](#)



Use Galaxy

[Main Server](#) ([Use](#))
[Other Servers](#) •
[Share](#) • [Search](#)

Communicate

[Support](#) • [News](#)
[Events](#) • [Twitter](#)
[Mailing Lists](#) (see [list](#))

Deploy Galaxy

[Get Galaxy](#) • [Cloud](#)
[Admin](#) • [Tool Co](#)
[Tool Shed](#) • [Sea](#)

Contribute

[Tool Shed](#) • [Sha](#)
[Issues & Reques](#)
[Support](#)

← → ↺ 131.156.41.196:8080/root

http://131.156.41.196:8080

☆ 📄 ☰

Galaxy

Analyze Data Workflow Shared Data Visualization Help User

Using 0 bytes

Tools

search tools

Get Data

Send Data

ENCODE Tools

Lift-Over

Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

Fetch Alignments

Get Genomic Scores

Operate on Genomic Intervals

Statistics

Wavelet Analysis

Graph/Display Data

Regional Variation

Multiple regression

Multivariate Analysis

Evolution

Motif Tools

Multiple Alignments

Metagenomic analyses

FASTA manipulation

NGS: QC and manipulation

NGS: Mapping

NGS: Indel Analysis

NGS: RNA Analysis

NGS: SAM Tools

NGS: GATK Tools (beta)

NGS: Peak Calling

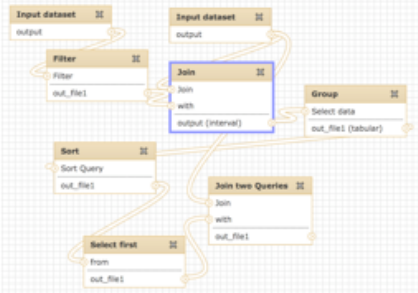
NGS: Simulation

✓ Hello world! It's running...

To customize this page edit static/welcome.html

WWFSMD?

grow noodly appendages...



usegalaxy.org

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

Unnamed history

0 bytes

ⓘ Your history is empty. Click 'Get Data' on the left pane to start

http://cys.bios.niu.edu:8080

11

Data intensive biology *for everyone.*

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.

Use Galaxy



Use the free public server

Get Galaxy



Install locally or in the cloud

Learn Galaxy



Screencasts, Galaxy 101, ...

Get Involved



Mailing lists, Tool Shed, wiki

[Search all resources](#)

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.

Galaxy Screencasts and Demos

Contents

1. [Getting Started](#)
2. [Using Galaxy 2012](#)
3. [Tool tutorials](#)
4. [Interval Operations tutorial](#)
5. [Examples of other analyses](#)
6. [Sample Tracking](#)
7. [Developers How To](#)
8. [Archives](#)

Screencasts are one of several ways to [learn](#) how to use Galaxy. These video tutorials cover many aspects of Galaxy, from simple tasks like uploading data, to complex analysis.

We are currently in the process of revamping Galaxy's screencast library. The current screencasts use a variety of technologies, including QuickTime (you may need to download the [QuickTime player](#)).

We hope to have both the screencast content and the screencast technology brought up to date by sometime

in 2012.

Learn

[Screencasts](#)[FAQ](#)[Interval Ops](#)[Datasets](#)[Pages](#)[Share](#)[FTP Upload](#)[Accounts](#)[Support](#)[Security](#)[Search](#)

Getting Started

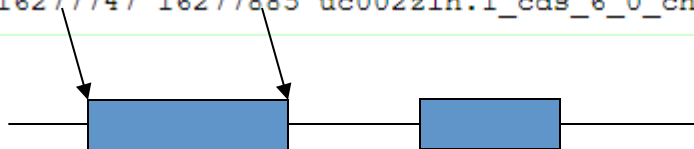
- [Galaxy 101](#)

What is the video about

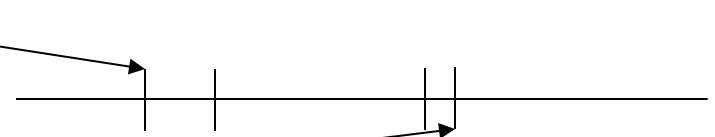
Find exons containing the largest number of SNPs in human chr22:

1. Download the exon data from UCSC
2. Download the SNP data from UCSC
3. Join the two files according to their chromosome locations
4. Count how many SNPs each exon has and generate a new file
5. Sort the file according to the number of SNPs in descending order
6. Check the top lines in the file

1.Chrom	2.Start	3.End	4.Name	5	6.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	-
chr22	16269872	16269943	uc002zlh.1_cds_4_0_chr22_16269873_r	0	-
chr22	16275206	16275277	uc002zlh.1_cds_5_0_chr22_16275207_r	0	-
chr22	16277747	16277885	uc002zlh.1_cds_6_0_chr22_16277748_r	0	-



1.Chrom	2.Start	3.End	4.Name	5	6.Strand
chr22	16050115	16050116	rs77005907	0	
chr22	16050251	16050252	rs3016036	0	+
chr22	16050352	16050353	rs56342815	0	+
chr22	16050352	16050353	rs2334386	0	+
chr22	16050374	16050375	rs2844882	0	+
chr22	16050407	16050408	rs2844883	0	+



1.Chrom	2.Start	3.End	4.Name	5	6.Strand	7	8	9	10	11	12				
chr22	16258185	16258303	uc002zlh.1_cds_1_0	chr22	16258186	r	0	-		chr22	16258278	16258279	rs2845178	0	+

Dataset 1

ctg15	10	49	Feature1
ctg15	70	119	Feature2
ctg15	170	209	Feature3
ctg15	180	229	Feature4

Dataset 2

ctg15	80	109	FeatureA
ctg15	150	199	FeatureB
ctg15	250	289	FeatureC
ctg15	270	309	FeatureD

Only records that are joined (INNER JOIN)

ctg15	70	119	Feature2	ctg15	80	109	FeatureA
ctg15	170	209	Feature3	ctg15	150	199	FeatureB
ctg15	180	229	Feature4	ctg15	150	199	FeatureB

All records of first dataset

ctg15	10	49	Feature1
ctg15	70	119	Feature2	ctg15	80	109	FeatureA
ctg15	170	209	Feature3	ctg15	150	199	FeatureB
ctg15	180	229	Feature4	ctg15	150	199	FeatureB

All records of second dataset

ctg15	70	119	Feature2	ctg15	80	109	FeatureA
ctg15	170	209	Feature3	ctg15	150	199	FeatureB
ctg15	180	229	Feature4	ctg15	150	199	FeatureB
.	.	.	.	ctg15	250	289	FeatureC
.	.	.	.	ctg15	270	309	FeatureD

All records of both datasets

ctg15	10	49	Feature1
ctg15	70	119	Feature2	ctg15	80	109	FeatureA
ctg15	170	209	Feature3	ctg15	150	199	FeatureB
ctg15	180	229	Feature4	ctg15	150	199	FeatureB
.	.	.	.	ctg15	250	289	FeatureC
.	.	.	.	ctg15	270	309	FeatureD

Register an account

The screenshot displays the Galaxy web interface. At the top, a dark navigation bar contains the 'Galaxy' logo and several menu items: 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The 'User' menu is highlighted with a red rectangle, and its dropdown menu is open, showing 'Login' and 'Register' options. Below the navigation bar, a green notification box on the left says 'Hello world! It's running...' with a checkmark icon and a link to 'static/welcome.html'. The left sidebar lists various tool categories: 'Get Data', 'Send Data', 'ENCODE Tools', 'Lift-Over', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Convert Formats', 'Extract Features', 'Fetch Sequences', 'Fetch Alignments', 'Get Genomic Scores', 'Operate on Genomic Intervals', 'Statistics', 'Wavelet Analysis', 'Graph/Display Data', 'Regional Variation', 'Multiple regression', 'Multivariate Analysis', and 'Evolution'. The main content area features a large graphic titled 'WWFSMD? grow noodly appendages...' with a diagram of interconnected workflow tools like 'Filter', 'Join', 'Group', 'Sort', 'Join two Queries', and 'Select first'. Below the graphic is the text 'usegalaxy.org'. At the bottom, a paragraph describes Galaxy as an open, web-based platform for data intensive biomedical research, mentioning its affiliation with Penn State and Emory University, and its support by NHGRI-NSF. The bottom status bar shows the URL 'cys.bios.niu.edu:8080/user/create?cntrlr=user'.

Galaxy Analyze Data Workflow Shared Data Visualization Help **User** Login Register Using 0 bytes

Tools

search tools

Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Wavelet Analysis
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution

Hello world! It's running...
To customize this page edit `static/welcome.html`


WWFSMD?
grow noodly appendages...

usegalaxy.org

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. Project is supported in part by NHGRI-NSF. The Huck


0 bytes
Your history is empty. Click 'Get Data' on the left pane to start

cys.bios.niu.edu:8080/user/create?cntrlr=user

 **Galaxy**

Analyze DataWorkflowShared Data▼Visualization▼Help▼User▼

Tools



[Get Data](#)
[Send Data](#)
[ENCODE Tools](#)
[Lift-Over](#)
[Text Manipulation](#)
[Filter and Sort](#)
[Join, Subtract and Group](#)
[Convert Formats](#)
[Extract Features](#)
[Fetch Sequences](#)
[Fetch Alignments](#)
[Get Genomic Scores](#)
[Operate on Genomic Intervals](#)
[Statistics](#)
[Wavelet Analysis](#)
[Graph/Display Data](#)
[Regional Variation](#)
[Multiple regression](#)
[Multivariate Analysis](#)
[Evolution](#)

Create account

Email address:


Password:

Confirm password:

Public name:

Your public name is an identifier that will be used to generate addresses for information you share publicly. Public names must be at least four characters in length and contain only lower-case letters, numbers, and the '-' character.

Submit

Histor
Unnan
0 byte
 You
Dat

Analyze DataWorkflowShared DataVisualizationHelpUser

Using 0 bytes

Now logged in as yyin@niu.edu.
Return to the home page.

History

HISTORY LISTS

Saved Histories

Histories Shared with Me

CURRENT HISTORY

Create New

Copy History

Copy Datasets

Share or Publish

Extract Workflow

Dataset Security

Resume Paused Jobs

Collapse Expanded Datasets

Include Deleted Datasets

Include Hidden Datasets

Unhide Hidden Datasets

Purge Deleted Datasets

Show Structure

Export to File

Delete

Delete Permanently

OTHER ACTIONS

WorkflowShared Data▼Visualization▼Help▼User▼Using 0 bytes

ed in as yyin@niu.edu.
to the home page.

History

Unnamed history

0 bytes

Your history is empty. Click 'Get Data' on the left pane to start

Change to galaxy-1

Step 1: get exon position data from UCSC

The screenshot displays the Galaxy web interface. At the top, the 'Galaxy' logo is on the left, and navigation tabs for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User' are in the center. The top right corner shows 'Using 0 bytes'. A green notification bar at the top center states: 'Now logged in as yyin@niu.edu. Return to the home page.' The left sidebar contains a 'Tools' section with a search bar and a list of tools. The 'Get Data' tool is highlighted with a red box, and the 'UCSC Main table browser' tool is also highlighted with a red box. The right sidebar shows a 'History' section with a message: 'Your history is empty. Click 'Get Data' on the left pane to start'.

Galaxy Analyze Data Workflow Shared Data Visualization Help User Using 0 bytes

Tools

search tools

Get Data

- Upload File from your computer
- UCSC Main table browser**
- UCSC Test table browser
- UCSC Archaea table browser
- BX table browser
- EBI SRA ENA SRA
- Get Microbial Data
- BioMart Central server
- BioMart Test server
- CBI Rice Mart rice mart
- GrameneMart Central server
- modENCODE fly server
- Elymine server
- Elymine test server

Now logged in as yyin@niu.edu.
[Return to the home page.](#)

History

galaxy-1
0 bytes

i Your history is empty. Click 'Get Data' on the left pane to start

We are connected to UCSC genome browser

We are going to download the exon position data

Human -> knownGene -> chr22

Galaxy Analyze Data Workflow Shared Data Visualization Help User Using 0 bytes

data to [GREAT](#). Refer to the [Credits](#) page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the [Sequence and Annotation Downloads](#) page.

clade: **genome:** **assembly:**

group: **track:** [add custom tracks](#)

[track hubs](#)

table: [describe table schema](#)

region: ☐ genome ☐ ENCODE Pilot regions ☒ position [lookup](#)

[define regions](#)

identifiers (names/accessions): [paste list](#) [upload list](#)

filter: [create](#)

intersection: [create](#)

correlation: [create](#)

output format: Send output to ☒ [Galaxy](#) ☐ [GREAT](#)

output file: (leave blank to keep output in browser)

file type returned: ☒ plain text ☐ gzip compressed

[get output](#) [summary/statistics](#)

To reset all user cart settings (including custom tracks), [click here](#).

Google ucsc bed format

BED format

[Index](#) ▶

BED format provides a flexible way to define the data lines that are displayed in an annotation track. BED lines have three required fields and nine additional optional fields. The number of fields per line must be consistent throughout any single set of data in an annotation track. The order of the optional fields is binding: lower-numbered fields must always be populated if higher-numbered fields are used.










If your data set is BED-like, but it is very large and you would like to keep it on your own server, you should use the [bigBed](#) data format.

The first three required BED fields are:

1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).
2. **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The *chromEnd* base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as *chromStart*=0, *chromEnd*=100, and span the bases numbered 0-99.

The 9 additional optional BED fields are:

4. **name** - Defines the name of the BED line. This label is displayed to the left of the BED line in the Genome Browser window when the track is open to full display mode or directly to the left of the item in pack mode.
5. **score** - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation data set, the *score* value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). This table shows the Genome Browser's translation of BED score values into shades of gray:

shade									
score in range	≤ 166	167-277	278-388	389-499	500-611	612-722	723-833	834-944	≥ 945

6. **strand** - Defines the strand - either '+' or '-'.
7. **thickStart** - The starting position at which the feature is drawn thickly (for example, the start codon in gene displays).
8. **thickEnd** - The ending position at which the feature is drawn thickly (for example, the stop codon in gene displays).
9. **itemRgb** - An RGB value of the form R,G,B (e.g. 255,0,0). If the track line *itemRgb* attribute is set to "On", this RGB value will determine the display color of the data contained in this BED line. NOTE: It is recommended that a simple color scheme (eight colors or less) be used with this attribute to avoid overwhelming the color resources of the Genome Browser and your Internet browser.
10. **blockCount** - The number of blocks (exons) in the BED line.

Choose **coding exons**

ble
a table
user
SRA
Data
al
server
rice
Central

Create one BED record per:

- ☐ Whole Gene
- ☐ Upstream by bases
- ☐ Exons plus bases at each end
- ☐ Introns plus bases at each end
- ☐ 5' UTR Exons
- ☒ **Coding Exons**
- ☐ 3' UTR Exons
- ☐ Downstream by bases

Note: if a feature is close to the beginning or end of a
are added, they may be truncated in order to avoid ex

Galaxy

Analyze Data

Workflow

Shared Data

Visualization

Help

User

Using 0 bytes

Tools

Get Data

Upload File from your computer

UCSC Main table browser

UCSC Test table browser

UCSC Archaea table browser

BX table browser

FBI SRA ENA SRA

Get Microbial Data

BioMart Central server

BioMart Test server

CBI Rice Mart rice mart

GrameneMart Central server

✓

The following job has been successfully added to the queue:
1: UCSC Main on Human: knownGene (chr22:1-51304566)
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.

galaxy-1

774.2 KB

1: UCSC Main on Human: knownGene (chr22:1-51304566)

History

galaxy-1

0 bytes

1: UCSC Main on Human: knownGene (chr22:1-51304566)

25

Tool tab

Browse tab

History tab

Analyze Data	Workflow	Shared Data ▾	Visualization ▾	Help ▾	User ▾	Using 774.2 KB
--------------	----------	---------------	-----------------	--------	--------	----------------

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
chr22	16269872	16269943	uc002zlh.1_cds_4_0_chr22_16269873_r	0
chr22	16275206	16275277	uc002zlh.1_cds_5_0_chr22_16275207_r	0
chr22	16277747	16277885	uc002zlh.1_cds_6_0_chr22_16277748_r	0
chr22	16279194	16279301	uc002zlh.1_cds_7_0_chr22_16279195_r	0
chr22	16280333	16280411	uc002zlh.1_cds_8_0_chr22_16280334_r	0
chr22	16258185	16258303	uc010gqp.2_cds_1_0_chr22_16258186_r	0
chr22	16266928	16267095	uc010gqp.2_cds_2_0_chr22_16266929_r	0
chr22	16268136	16268181	uc010gqp.2_cds_3_0_chr22_1	
chr22	16269872	16269943	uc010gqp.2_cds_4_0_chr22_1	
chr22	16275206	16275277	uc010gqp.2_cds_5_0_chr22_1	
chr22	16277747	16277885	uc010gqp.2_cds_6_0_chr22_1	
chr22	16279194	16279301	uc010gqp.2_cds_7_0_chr22_1	
chr22	16282144	16282318	uc010gqp.2_cds_8_0_chr22_1	
chr22	16282477	16282592	uc010gqp.2_cds_9_0_chr22_1	
chr22	16287253	16287885	uc010gqp.2_cds_10_0_chr22_1	
chr22	16266930	16267095	uc002zlj.1_cds_0_0_chr22_16	
chr22	16268136	16268181	uc002zlj.1_cds_1_0_chr22_16	
chr22	16269872	16269943	uc002zlj.1_cds_2_0_chr22_16	
chr22	16275206	16275277	uc002zlj.1_cds_3_0_chr22_16	
chr22	16277747	16277885	uc002zlj.1_cds_4_0_chr22_16	

History

galaxy-1

774.2 KB

1: UCSC Main on Human:
knownGene (chr22:1-
51304566)



Display data in browser

1: UCSC Main on Human: knownGene
(chr22:1-51304566)



12,333 regions
format: bed, database: hg19



display in IGB [Local](#) [Web](#)

display at Ensembl [Current](#)

display at RViewer [main](#)



1.Chrom	2.Start	3.End	4.Name
chr22	16258185	16258303	uc002zlh.1_cds_1_0_chr22
chr22	16266928	16267095	uc002zlh.1_cds_2_0_chr22
chr22	16268136	16268181	uc002zlh.1_cds_3_0_chr22
chr22	16269872	16269943	uc002zlh.1_cds_4_0_chr22
chr22	16275206	16275277	uc002zlh.1_cds_5_0_chr22
chr22	16277747	16277885	uc002zlh.1_cds_6_0_chr22

BED format

Now download **SNP position data**

Tools

Get Data

- Upload File from your computer
- UCSC Main table browser**
- UCSC Test table browser
- UCSC Archaea table browser
- BX table browser
- EBI SRA ENA SRA
- Get Microbial Data
- BioMart Central server
- BioMart Test server
- CBI Rice Mart rice mart
- GrameneMart Central server
- modENCODE fly

with these data. All tables can be downloaded in their entirety from the [Sequence Downloads](#) page.

clade: Mammal **genome:** Human **assembly:** Feb. 2009 (GRC38)

group: Variation and Repeats **track:** SNPs (131)

table: snp131 [describe table schema](#)

region: ☐ genome ☐ ENCODE Pilot regions ☒ position chr22

[define regions](#)

identifiers (names/accessions): [paste list](#) [upload list](#)

filter: [create](#)

intersection: [create](#)

correlation: [create](#)

output format: BED - browser extensible data ☒ Send output to [Google](#)

output file: (leave blank to keep output in browser)

file type returned: ☒ plain text ☐ gzip compressed

[get output](#) [summary/statistics](#)

To reset all user cart settings (including custom tracks), [click here](#).

our

Output snp131 as BED

☐ Include [custom track](#) header:

name=

description=

visibility= ▼

url=

Create one BED record per:

☒ Whole Gene

☐ Upstream by bases

☐ Downstream by bases

Note: if a feature is close to the beginning or end of a chromosome, they may be truncated in order to avoid extending beyond the chromosome boundaries.

le

er

tral

Analyze DataWorkflowShared DataVisualizationHelpUserUsing 14.7 MB

AttributesConvert FormatDatatypePermissions

Edit Attributes

Name:

UCSC Main on Human: snp131 (chr22)

Info:

Annotation / Notes:

None

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build:

Human Feb. 2009 (GRCh37/hg19) (hg19)

Number of comment lines:

☐

Chrom column:

History

galaxy-1
14.7 MB

2: UCSC Main on Human: snp131
(chr22:1-51304566)

1: UCSC Main on Human: knownGene
(chr22:1-51304566)

Change data set names

Type in
SNPs
and then press Enter

Similarly for 1, change to
Exons

✓ Attributes updated

Attributes

Convert Format

Datatype

Permissions

Edit Attributes

Name:

Exons

Info:

Annotation / Notes:

None

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build:

Human Feb. 2009 (GRCh37/hg19) (hg19)

Number of comment lines:

History

galaxy-1

14.7 MB

2: SNPs

1: Exons



Galaxy

Analyze Data

Workflow

Help ▾

User ▾

Using 14.7 MB

chr22	16050115	16050116	rs77005907	0	-		
chr22	16050251	16050252	rs3016036	0	+		
chr22	16050352	16050353	rs56342815	0	+		
chr22	16050352	16050353	rs2334386	0	+		
chr22	16050374	16050375	rs2844882	0	+		
chr22	16050407	16050408	rs2844883	0	+		
chr22	16050454	16050455	rs61969399	0	-		
chr22	16050611	16050612	rs2186463	0	+		
chr22	16050677	16050678	rs2186465	0	+		
chr22	16050713	16050714	rs61969398	0	-		
chr22	16050821	16050822	rs12172168	0	+		
chr22	16050966	chr22	16258185	16258303	uc002zlh.1_cds_1_0_chr22_16258186_r	0	-
chr22	16050993	chr22	16266928	16267095	uc002zlh.1_cds_2_0_chr22_16266929_r	0	-
chr22	16050993	chr22	16268136	16268181	uc002zlh.1_cds_3_0_chr22_16268137_r	0	-
chr22	16051106	chr22	16269872	16269943	uc002zlh.1_cds_4_0_chr22_16269873_r	0	-
chr22	16051106	chr22	16275206	16275277	uc002zlh.1_cds_5_0_chr22_16275207_r	0	-
chr22	16051208	chr22	16277747	16277885	uc002zlh.1_cds_6_0_chr22_16277748_r	0	-
chr22	16051240	chr22	16279194	16279301	uc002zlh.1_cds_7_0_chr22_16279195_r	0	-
chr22	16051248	chr22	16280333	16280411	uc002zlh.1_cds_8_0_chr22_16280334_r	0	-
chr22	16051254	chr22	16258185	16258303	uc010gqp.2_cds_1_0_chr22_16258186_r	0	-
chr22	16051294	chr22	16266928	16267095	uc010gqp.2_cds_2_0_chr22_16266929_r	0	-
chr22	16051294	chr22	16268136	16268181	uc010gqp.2_cds_3_0_chr22_16268137_r	0	-
chr22	16051246	chr22	16269872	16269943	uc010gqp.2_cds_4_0_chr22_16269873_r	0	-
		chr22	16275206	16275277	uc010gqp.2_cds_5_0_chr22_16275207_r	0	-
		chr22	16277747	16277885	uc010gqp.2_cds_6_0_chr22_16277748_r	0	-
chr22	16279194	chr22	16279301	uc010gqp.2_cds_7_0_chr22_16279195_r	0	-	

History

galaxy-1

14.7 MB

2: SNPs

1: Exons

Tools

Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Wavelet Analysis
Graph/Display Data
Regional Variation
Multiple regression

Attributes updated

Attributes

Edit Attributes

Name:
Exons

Info:

Annotation:
None

Add an annotation when a history item is created

Database/Build:
Human Feb. 2009

Number of comment lines:
☐

Tools

Operate on Genomic Intervals

- Intersect the intervals of two datasets
- Subtract the intervals of two datasets
- Merge the overlapping intervals of a dataset
- Concatenate two datasets into one dataset
- Base Coverage of all intervals
- Coverage of a set of intervals on second set of intervals
- Complement intervals of a dataset
- Cluster the intervals of a dataset
- **Join** the intervals of two datasets side-by-side
- Get flanks returns flanking region/s for every gene

History

galaxy-1
14.7 MB

2: SNPs

1: Exons

Analyze DataWorkflowShared Data▼Visualization▼Help▼User▼

Using 14.7 MB

Join (version 1.0.0)

Join:

1: Exons

First dataset

with:

2: SNPs

Second dataset

with min overlap:

1

(bp)

Return:

Only records that are joined (INNER JOIN)

Execute

TIP: If your dataset does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.

Screenscasts!

See Galaxy Interval Operation Screenscasts (right click to open this link in another

History

galaxy-1

14.7 MB

2: SNPs

1: Exons






The following job has been successfully added to the queue:

3: Join on data 2 and data 1





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.

Click on the link to expand the panel

galaxy-1
16.2 MB

3: Join on data 2 and data 1   

15,469 regions
format: interval, database: hg19



display at Ensembl [Current](#)
display at RViewer [main](#)





1.Chrom	2.Start	3.End	4.Name
chr22	16258185	16258303	uc002z1h.1_c
chr22	16266928	16267095	uc002z1h.1_c
chr22	16266928	16267095	uc002z1h.1_c
chr22	16266928	16267095	uc002z1h.1_c
chr22	16266928	16267095	uc002z1h.1_c
chr22	16266928	16267095	uc002z1h.1_c
chr22	16269872	16269943	uc002z1h.1_c




Becomes green after it is done




History  


galaxy-1

14.7 MB  


 **3: Join on data 2 and data 1**   

2: SNPs   




1: Exons   




History  

galaxy-1

16.2 MB  

3: Joined   

2: SNPs   

1: Exons   

Change the name to **Joined**

4 SNPs are located in exon uc002zlh.1_cds_2_0_chr22_16266929_r

Now we need to count for each exon, in how many lines do they appear

Or

more specifically, count how many times each word in column 4 appear if we group them

chr22	16258185	16258303	uc002zlh.1_cds_1_0_chr22_16258186_r	0	-	chr22	16258278	16258279	rs2845178	0	+
chr22	16266928	16267095	uc002zlh.1_cds_2_0_chr22_16266929_r	0	-	chr22	16267037	16267038	rs2818572	0	+
chr22	16266928	16267095	uc002zlh.1_cds_2_0_chr22_16266929_r	0	-	chr22	16267031	16267032	rs7292200	0	+
chr22	16266928	16267095	uc002zlh.1_cds_2_0_chr22_16266929_r	0	-	chr22	16266963	16266964	rs10154680	0	+
chr22	16266928	16267095	uc002zlh.1_cds_2_0_chr22_16266929_r	0	-	chr22	16267011	16267012	rs7290262	0	+
chr22	16269872	16269943	uc002zlh.1_cds_4_0_chr22_16269873_r	0	-	chr22	16269933	16269934	rs2845206	0	+
chr22	16275206	16275277	uc002zlh.1_cds_5_0_chr22_16275207_r	0	-	chr22	16275252	16275253	rs8142076	0	+
chr22	16275206	16275277	uc002zlh.1_cds_5_0_chr22_16275207_r	0	-	chr22	16275237	16275238	rs2845214	0	+
chr22	16277747	16277885	uc002zlh.1_cds_6_0_chr22_16277748_r	0	-	chr22	16277756	16277757	rs79385954	0	+
chr22	16277747	16277885	uc002zlh.1_cds_6_0_chr22_16277748_r	0	-	chr22	16277756	16277757	rs2845218	0	+
chr22	16277747	16277885	uc002zlh.1_cds_6_0_chr22_16277748_r	0	-	chr22	16277851	16277852	rs11489067	0	+
chr22	16277747	16277885	uc002zlh.1_cds_6_0_chr22_16277748_r	0	-	chr22	16277818	16277819	rs2073406	0	+
chr22	16277747	16277885	uc002zlh.1_cds_6_0_chr22_16277748_r	0	-	chr22	16277879	16277880	rs8135863	0	+
chr22	16279194	16279301	uc002zlh.1_cds_7_0_chr22_16279195_r	0	-	chr22	16279241	16279242	rs56237058	0	+
chr22	16279194	16279301	uc002zlh.1_cds_7_0_chr22_16279195_r	0	-	chr22	16279241	16279242	rs3000542	0	+

Exon data

SNP data

History

galaxy-1

16.2 MB

3: Joined

2: SNPs

1: Exons

Galaxy Analyze Data Workflow Shared Data

Tools

[Send Data](#)
[ENCODE Tools](#)
[Lift-Over](#)
[Text Manipulation](#)
[Filter and Sort](#)
1. [Join, Subtract and Group](#)
▪ [Join two Datasets](#) side by side on a specified field
▪ [Compare two Datasets](#) to find common or distinct rows
▪ [Subtract Whole Dataset](#) from another dataset
2. [Group data by a column and perform aggregate operation on other columns.](#)
▪ [Column Join](#)
[Convert Formats](#)
[Extract Features](#)
[Fetch Sequences](#)
[Fetch Alignments](#)
[Get Genomic Scores](#)
[Operate on Genomic Intervals](#)

Group (version 2.0.0)

Select data:
3. [3: Joined](#)
Dataset missing? See TIP below.

Group by column:
[c4](#)

Ignore case while grouping?:
☐

Operations
4. Click on add new operation

Operation 1

Type:
5. [Count](#)

On column:
6. [c4](#)

Round result to nearest integer?:
[NO](#)

[Remove Operation 1](#)

[Add new Operation](#)

7. [Execute](#)

Now for each exon we know how many times they appeared in the previous “Joined” data file, i.e. the SNP numbers. However, we need to find out which one has the largest number in col 2. Note there are 6565 lines in the file

The screenshot shows the Galaxy web interface. At the top, there are navigation tabs: 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The top right corner indicates 'Using 16.2 MB'.

On the left, a list of genomic data files is displayed, each with a file ID and a count in the 4th column. For example, 'uc002zlh.1_cds_1_0_chr22_16258186_r' has a count of 1.

In the center, a tool output window for '5: Grouped 4th col' is open. It shows '6,565 lines' and the command 'format: tabular, database: hg19 --Group by c4: count[c4]'. Below the command, a preview of the output is shown, starting with '1' and followed by the same genomic file IDs and counts as in the left panel.

On the right, the 'History' panel shows a list of workflow items. The items are 'galaxy-1' (16.4 MB), '5: Group on data 3', '3: Joined', '2: SNPs', and '1: Exons'. The '5: Group on data 3' item is highlighted with a red box, and a red arrow points from it to the '5: Grouped 4th col' tool output window. Another red arrow points from the '5: Grouped 4th col' tool output window to the '5: Grouped 4th col' item in the history panel.

At the bottom right, the page number '37' is visible.

Galaxy

Tools

[Get Data](#)
[Send Data](#)
[ENCODE Tools](#)
[Lift-Over](#)
[Text Manipulation](#)
[Filter and Sort](#)

- [Filter](#) data on any column using simple expressions
- [Sort](#) data in ascending or descending order
- [Select](#) lines that match an expression

GFF

- [Extract features](#) from GFF data
- [Filter GFF data by attribute](#) using simple expressions
- [Filter GFF data by feature](#) count using simple expressions

Sort (version 1.0.1)

Sort Dataset:
5: Grouped 4th col

on column:
c2

with flavor:
Numerical sort

everything in:
Descending order

Column selections
Add new Column selection

Execute

How do we pick out the top 5?

Analyze Data Workflow Shared Data Visualization Help User Using 16.7 MB

uc010gsw.2_cds_1_0_chr22_21480537_r	67
uc021wmb.1_cds_0_0_chr22_21480537_r	67
uc002zoc.3_cds_0_0_chr22_18834445_f	58
uc021wnd.1_cds_0_0_chr22_24647973_f	50
uc021wmc.1_cds_0_0_chr22_21637809_f	47
uc003bhh.3_cds_0_0_chr22_46652458_r	46
uc002zsd.4_cds_0_0_chr22_20456382_r	41
uc002zuq.4_cds_0_0_chr22_21738148_f	41
uc002zuy.4_cds_0_0_chr22_21900346_r	41
uc003atq.1_cds_12_0_chr22_38119192_f	37
uc003atr.3_cds_6_0_chr22_38119192_f	37
uc003ats.1_cds_4_0_chr22_38119192_f	37
uc003atu.3_cds_4_0_chr22_38119192_f	37
uc011aho.1_cds_0_0_chr22_20640678_r	36
uc003afo.3_cds_3_0_chr22_29884838_f	35
uc011agd.2_cds_0_0_chr22_16448824_r	35
uc003ara.3_cds_0_0_chr22_37602586_r	28

History

- galaxy-1
16.7 MB
- 6: Sort on data 5
- 5: Grouped 4th col
- 3: Joined
- 2: SNPs
- 1: Exons

6: Sorted

5: Grouped 4th col

3: Joined

2: SNPs

1: Exons

Tools

Lift-Over

Text Manipulation

- Add column to an existing dataset
- Compute an expression on every row
- Concatenate datasets tail-to-head
- Cut columns from a table
- Merge Columns together
- Convert delimiters to TAB
- Create single interval as a new dataset
- Change Case of selected columns
- Paste two files side by side
- Remove beginning of a file
- Select random lines from a file
- Select first lines from a dataset
- Select last lines from a dataset

Select first (version 1.0.0)

Select first:

5

lines

from:

6: Sorted

Execute



What it does



This tool outputs specified number of lines from the




How do we find back the exon position info?




Shared Data ▾ Visualization ▾ Help ▾ User ▾ Using 16.7 MB




uc010gsw.2_cds_1_0_chr22_21480537_r	67
uc021wmb.1_cds_0_0_chr22_21480537_r	67
uc002zoc.3_cds_0_0_chr22_18834445_f	58
uc021wnd.1_cds_0_0_chr22_24647973_f	50
uc021wmc.1_cds_0_0_chr22_21637809_f	47




History  




galaxy-1
16.7 MB  




7: Select first on data 6   



6: Sorted   




5: Grouped 4th col   




3: Joined   




2: SNPs   




1: Exons   




galaxy-1
16.7 MB  




7: top 5   

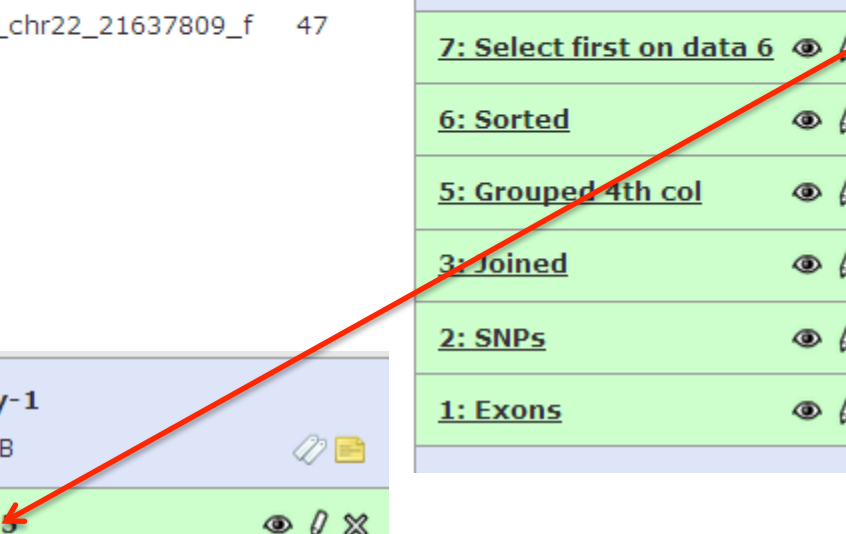
6: Sorted   

5: Grouped 4th col   

3: Joined   

2: SNPs   

1: Exons   



Tools

search tools

[Get Data](#)

[Send Data](#)

[ENCODE Tools](#)

[Lift-Over](#)

[Text Manipulation](#)

[Filter and Sort](#)

[Join, Subtract and Group](#)

Join two Datasets side by side on a specified field

Compare two Datasets to find common or distinct rows

Subtract Whole Dataset from another dataset

Group data by a column and perform aggregate operation on other columns.

Column Join

[Convert Formats](#)

[Extract Features](#)

[Fetch Genomes](#)

Join two Datasets (version 2.0.2)

Join:

7: top 5

using column:

c1

with:

1: Exons

and column:

c4

Keep lines of first input that do not join with second input:

No

Keep lines of first input that are incomplete:

No

Fill empty columns:

No

Execute

42

I do not need these columns

Galaxy

Analyze Data Workflow Shared Data Visualization Help User

Using 16.7 MB

uc010gsw.2_cds_1_0_chr22_21480537_r	67	chr22	21480536	21481925	uc010gsw.2_cds_1_0_chr22_21480537_r	0	-
uc021wmb.1_cds_0_0_chr22_21480537_r	67	chr22	21480536	21481925	uc021wmb.1_cds_0_0_chr22_21480537_r	0	-
uc002zoc.3_cds_0_0_chr22_18834445_f	58	chr22	18834444	18835833	uc002zoc.3_cds_0_0_chr22_18834445_f	0	+
uc021wnd.1_cds_0_0_chr22_24647973_f	50	chr22	24647972	24649256	uc021wnd.1_cds_0_0_chr22_24647973_f	0	+
uc021wmc.1_cds_0_0_chr22_21637809_f	47	chr22	21637808	21638558	uc021wmc.1_cds_0_0_chr22_21637809_f	0	+

History

galaxy-1

16.7 MB

8: top 5 joined exon

7: top 5

6: Sorted


5: Grouped 4th col

3: Joined

2: SNPs

1: Exons

1,2,3,4,5,8

 **Galaxy**

Analyze DataWorkflowShared Data▼Visualiza

Tools

Get Data

Send Data

ENCODE Tools

Lift-Over

Text Manipulation

- [Add column](#) to an existing dataset
- [Compute](#) an expression on every row
- [Concatenate datasets](#) tail-to-head
- **[Cut columns from a table](#)**
- [Merge Columns](#) together
- [Convert](#) delimiters to TAB
- [Create single interval](#) as a new dataset
- [Change Case](#) of selected columns

Cut (version 1.0.1)

Cut columns:

Delimited by:

From:

Execute

⚠ WARNING: This tool breaks column assignments. T assignments run the tools and click on the pencil icon

i The output of this tool is always in tabular format (e.g commas, they will be replaced with tabs). For example

Cutting columns 1 and 3 from:

```
apple,is,good
windows,is,bad
```


uc010gsw.2_cds_1_0_chr22_21480537_r	67	chr22	21480536	21481925	-
uc021wmb.1_cds_0_0_chr22_21480537_r	67	chr22	21480536	21481925	-
uc002zoc.3_cds_0_0_chr22_18834445_f	58	chr22	18834444	18835833	+
uc021wnd.1_cds_0_0_chr22_24647973_f	50	chr22	24647972	24649256	+
uc021wmc.1_cds_0_0_chr22_21637809_f	47	chr22	21637808	21638558	+

History  	
galaxy-1 16.7 MB  	
<u>9: top 5 cut</u>	  
<u>8: top 5 joined exon</u>	  
<u>7: top 5</u>	  
<u>6: Sorted</u>	  
<u>5: Grouped 4th col</u>	  
<u>3: Joined</u>	  
<u>2: SNPs</u>	  
<u>1: Exons</u>	  

Saved Histories

search history names and tags



[Advanced Search](#)

<input type="checkbox"/>	Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated ↑
<input type="checkbox"/>	galaxy-1 ▾	8	0 Tags		16.7 MB	~ 2 hours ago	16 minutes ago
<input type="checkbox"/>	Unnamed history ▾		0 Tags		0 bytes	~ 2 hours ago	~ 2 hours ago

For 0 selected histories:

Rename

Delete

Delete Permanently

Undelete

Histories that have been deleted for more than a time period specified by the Galaxy administrator(s) may be permanently deleted.

History



HISTORY LISTS

Saved Histories

Histories Shared with Me

CURRENT HISTORY

Create New

Copy History

Copy Datasets

Share or Publish

Extract Workflow

Dataset Security

Resume Paused Jobs

Collapse Expanded Datasets

Include Deleted Datasets

Include Hidden Datasets

Unhide Hidden Datasets

Purge Deleted Datasets

Show Structure

Export to File

Delete

Delete Permanently

OTHER ACTIONS

Import from File

The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow.

Tools which cannot be run interactively and thus cannot be incorporated into a workflow will be shown in gray.

Workflow name**Create Workflow**

Check all

Uncheck all

Tool**History items created**

UCSC Main

*This tool cannot be used in workflows***1: Exons**☒ Treat as input dataset

UCSC Main

*This tool cannot be used in workflows***2: SNPs**☒ Treat as input dataset

Join

☒ Include "Join" in workflow**3: Joined**

Group

History

HISTORY LISTS

Saved Histories

Histories Shared with Me

CURRENT HISTORY

Create New

Copy History

Copy Datasets

Share or Publish

Extract Workflow

Dataset Security

Resume Paused Jobs

Collapse Expanded Datasets

Include Deleted Datasets

Include Hidden Datasets

Unhide Hidden Datasets

Purge Deleted Datasets

Show Structure

Export to File

Delete

Delete Permanently

OTHER ACTIONS

Your workflows

Name

Workflow constructed from history 'galaxy-1' ▼

Workflows shared with others

No workflows have been shared

Other options

Configure your workflow menu

Edit

Run

Share or Publish

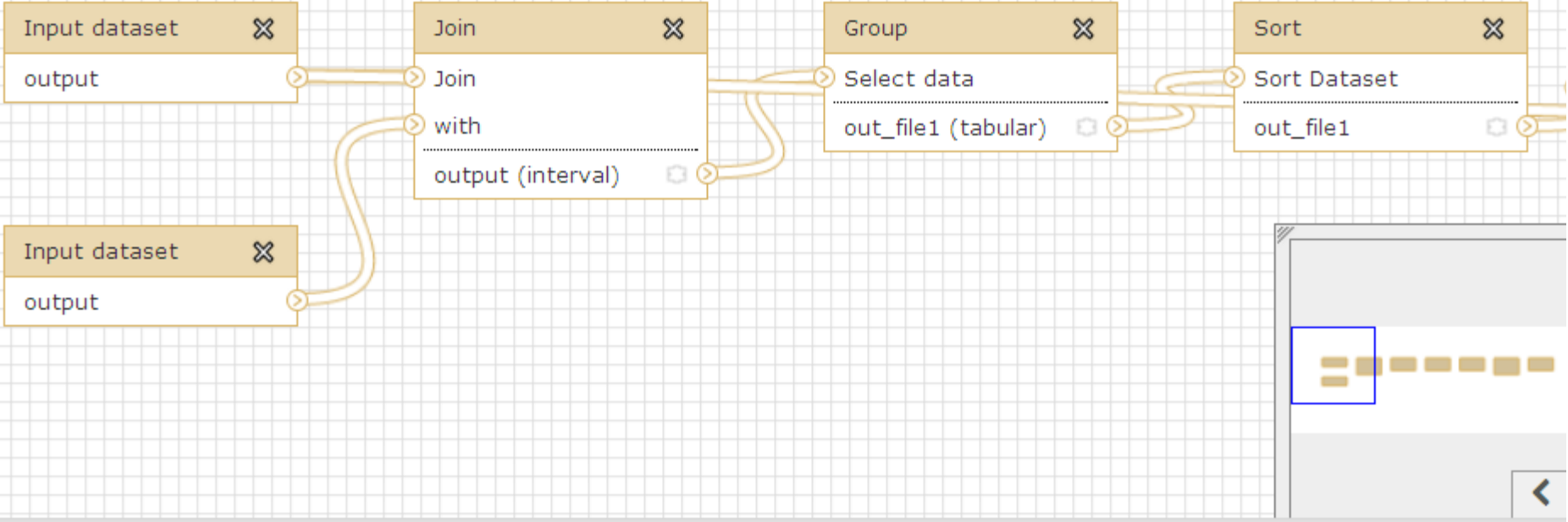
Download or Export

Copy

Rename

View

Delete



Re-use the workflow we just saved

Let's try to use All SNP 137 and all the rest remains the same

Next lecture: Galaxy II