Galaxy Workflows

Workflows Workshop

August 9-10, 2016
Online

Dave Clements & Enis Afgan
Galaxy Team
Johns Hopkins University
http://galaxyproject.org/

XSEDE
Extreme Science and Engineering Discovery Environment

Galaxy PROJECT

#usegalaxy @galaxyproject

Agenda

Launching Galaxy in Jetstream

A quick stroll through the Galaxy

Demonstrate Galaxy by addressing a specific question

Turning that analysis into a reusable workflow

Shutting down Galaxy in Jetstream

Galaxy Ecosystem

bit.ly/ww_gxy_slides

Launching Galaxy in Jetstream

bit.ly/gxyjets

(https://wiki.galaxyproject.org/Cloud/Jetstream)

Agenda

Launching Galaxy in Jetstream

A quick stroll through the Galaxy

Demonstrate Galaxy by addressing a specific question

Turning that analysis into a reusable workflow

Shutting down Galaxy in Jetstream

Galaxy Ecosystem

bit.ly/ww_gxy_slides

What is Galaxy?

Keith Bradnam's definition:

"A web-based platform that provides a simplified interface to many popular bioinformatics tools."

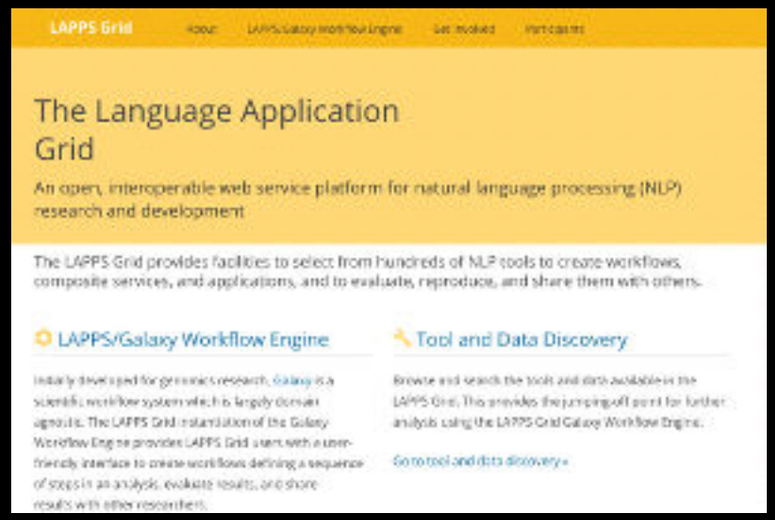
From

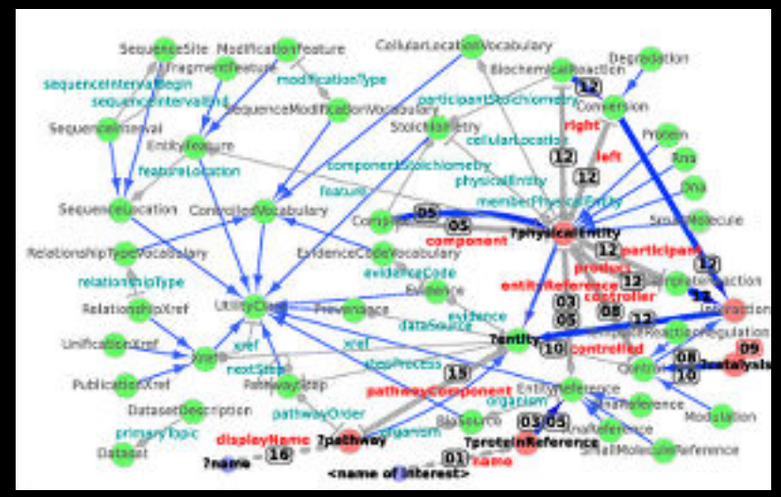
"13 Questions You May Have About Galaxy"

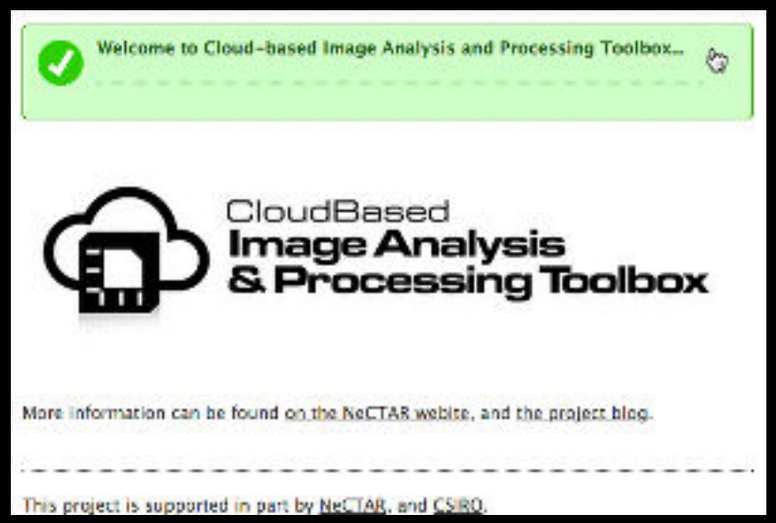
http://bit.ly/13questions

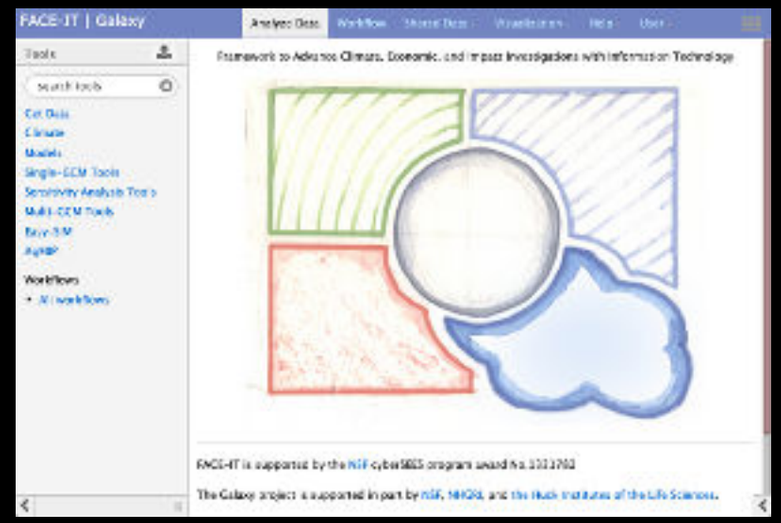
http://galaxyproject.org

Galaxy was created for genomics.
Now used across a spectrum of disciplines









Metabolomics, Natural Language, Image Analysis,
Climate Change, Social Science, Cosmology, ...

Galaxy is **available** several ways ...

bit.ly/ww_gxy_slides

As a free for everyone service on the web: usegalaxy.org

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 384.0 GB

We have updated Galaxy to the forthcoming 16.04 release and made some configuration and deployment changes, so you may encounter a few problems. We are working on these, and please feel free to report any errors you encounter.

Tools

search tools

- Get Data
- Send Data
- Lift-Over
- Text Manipulation
- Datamash
- Convert Formats
- Filter and Sort
- Join, Subtract and Group
- Fetch Alignments/Sequences
- NGS: QC and manipulation
- NGS: DeepTools
- NGS: Mapping
- NGS: RNA Analysis
- NGS: SAMtools
- NGS: BamTools
- NGS: Picard
- NGS: VCF Manipulation
- NGS: Peak Calling
- NGS: Variant Analysis
- NGS: RNA Structure
- NGS: Du Novo
- NGS: Gemini
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- CloudMap
- Phenotype Association

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

Running Your Own

Understanding how Galaxy works

An in-depth tutorial

Tweets by @galaxyproject

Galaxy Project Retweeted

Hans-Rudolf Hotz @hrhotz
check out this useful resource for #usegalaxy training and teaching material from @bjoerngruening and his team: wiki.galaxyproject.org/Teach/Resource…

11h

Galaxy Project @galaxyproject
GCC2016 student/postdoc #underrepresentedgroups scholarships. Apply by May 1 bit.ly/gcc2016scholar… #usegalaxy

Embed View on Twitter

PENNSTATE 1855 | JOHNS HOPKINS UNIVERSITY | TACC | CYVERSE

Galaxy is available as Open Source Software

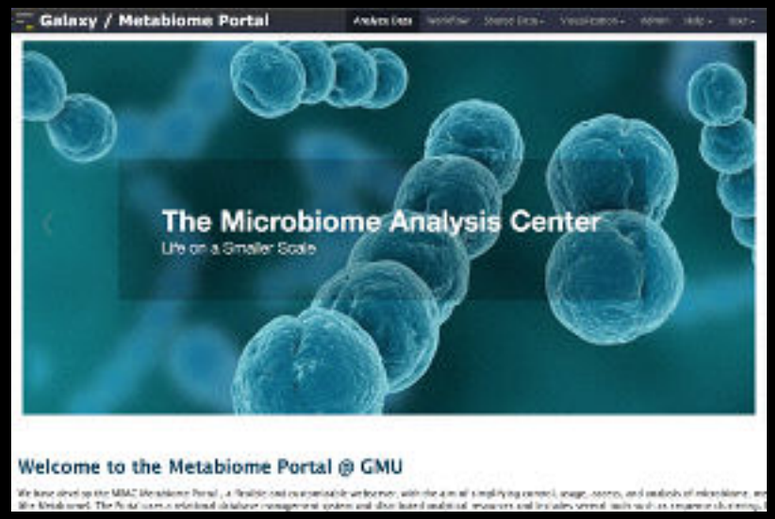
Galaxy is installed in locations around the world.

http://getgalaxy.org

Explore the Galaxy with

RNA-Rocket

PATHOGENPORTAL
THE BIOINFORMATICS RESOURCE CENTERS PORTAL



香港中文大學 - 華大基因跨組學創新研究院
CUHK-BGI Innovation Institute of Trans-Omics

(GIGA)^n Galaxy
by CBIIT

Integrated publishing of workflows from (GIGA)^n SCIENCE

Cistrome

A Galaxy Server dedicated to ChIP-* analysis

080+

Public Galaxy Servers
and *still* counting

The Genomic HyperBrowser

Powered by Galaxy

SCDE
STEM CELL DISCOVERY ENGINE

Experiments Connected

South Green
bioinformatics platform

Genomic analysis tools for southern and Mediterranean plants

bit.ly/gxyServers

Galaxy is available on the Cloud

http://aws.amazon.com/education
http://globus.org/
http://wiki.galaxyproject.org/Cloud

Via

We'll use this now

bit.ly/ww_gxy_slides

Agenda

Launching Galaxy in Jetstream

A quick stroll through the Galaxy

Demonstrate Galaxy by addressing a specific question

Turning that analysis into a reusable workflow

Shutting down Galaxy in Jetstream

Galaxy Ecosystem

bit.ly/ww_gxy_slides

Specific Question: Repeats in Genes?

Which genes physically overlap with large numbers of repeats?

bit.ly/ww_gxy_slides

Genes: Exons, Introns, and alternative versions

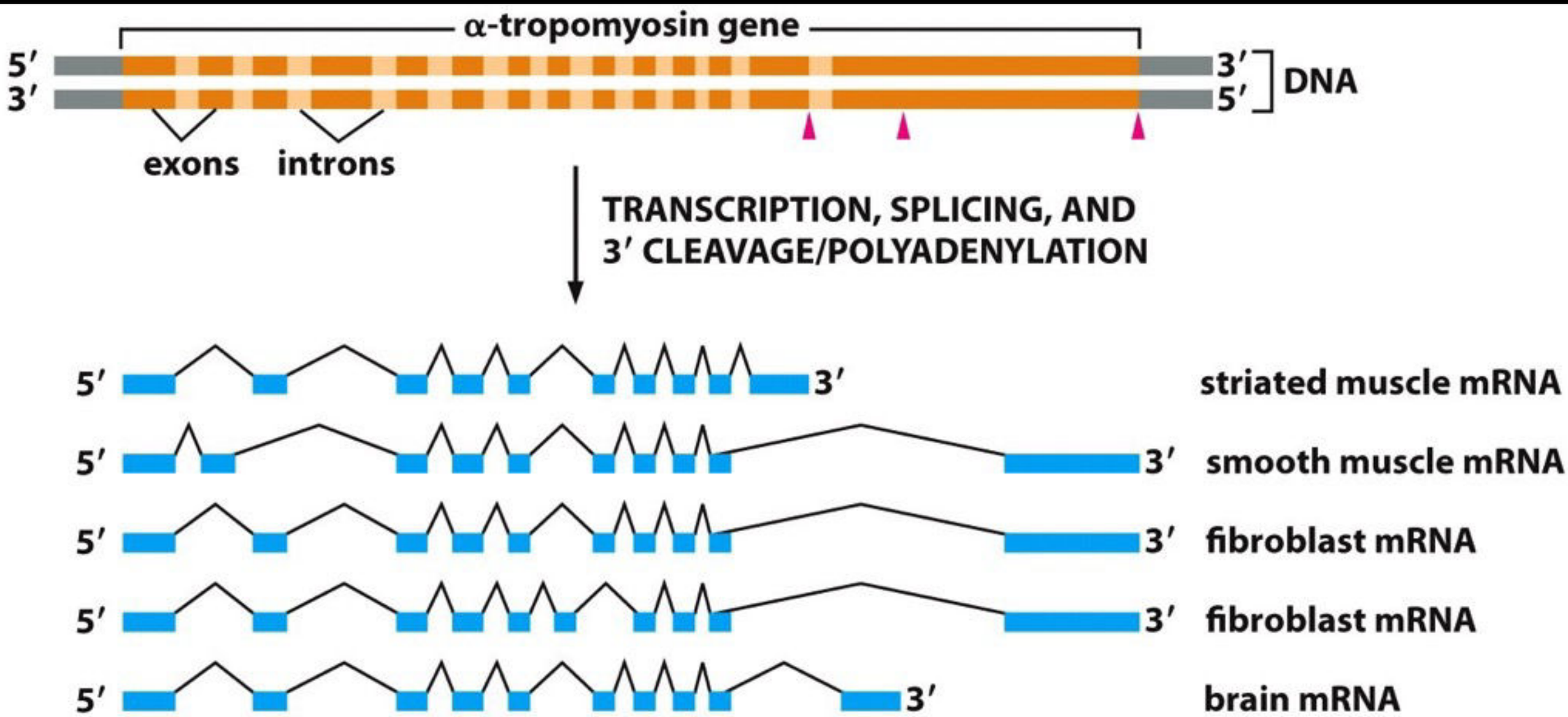


Figure 7-21 Essential Cell Biology 3/e (© Garland Science 2010)

http://oregonstate.edu/instruction/bi314/fall12/geneexpression.html

Genes: Exons, Introns, and alternative versions

What to remember:

Exons become **proteins** and **proteins** become **you**.

Exons can be **combined** to create **different proteins**.

Warning:

"Biology is a system of exceptions." Tom Conlin, 2000

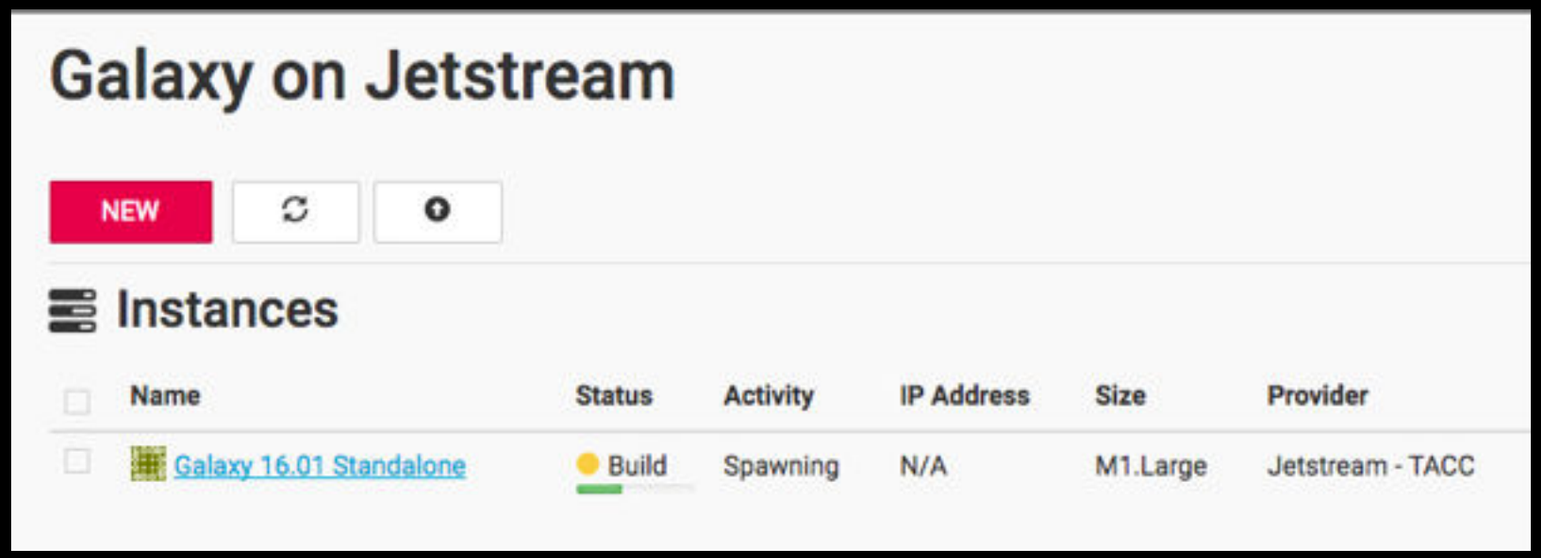
Repeats: Simple and Complex

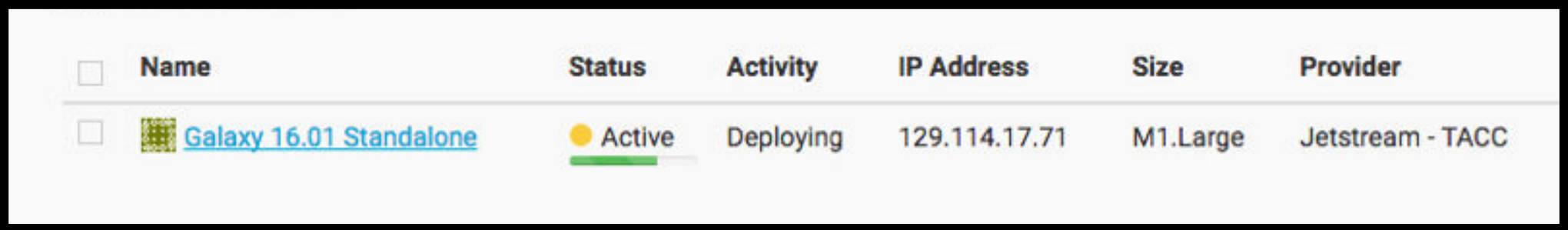
Simple repeats:
DNA **stuttering**

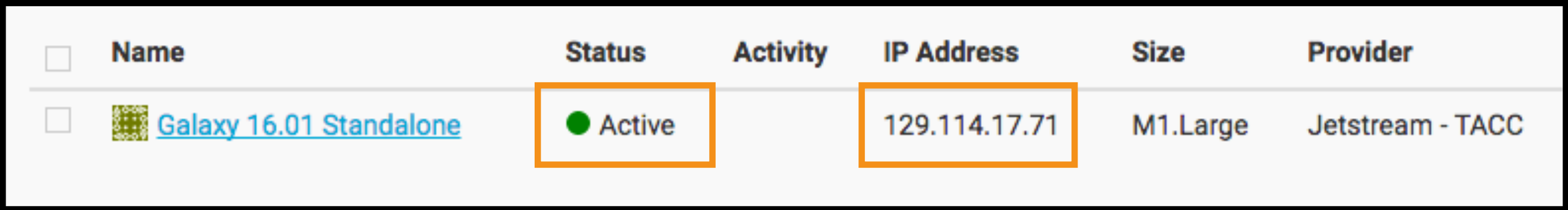
Complex repeats:
DNA that can **replicate** and **insert copies** of themselves

Half the human genome is **repetitive**
1/50th of the human genome is **exons**

Galaxy: Should be active:







Copy & paste IP address into a new browser tab.

Tools

search tools

Get Data
Lift-Over
Text Manipulation
Datamash
Convert Formats
Filter and Sort
Join, Subtract and Group
Fetch Sequences
Fetch Alignments
Extract Features
Statistics
Graph/Display Data
BEDTools
NGS: QC and manipulation
NGS: Mapping
NGS: RNA Analysis
NGS: SAMtools
NGS: BamTools
NGS: Picard
NGS: VCF Manipulation
NGS: Peak Calling
NGS: Variant Analysis
NGS: RNA Structure
NGS: Gemini

Welcome to Galaxy on the Jetstream Cloud

***Galaxy on the Jetstream Cloud* is ready for use!**

To learn how to use Galaxy please see the wiki.
To install new tools to your Galaxy follow the tutorial.

Thank you for using Galaxy.

Galaxy is an open, web-based platform for data intensive biomedical research. The Galaxy team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State, and the Department of Biology and Computer Science at Johns Hopkins 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 Johns Hopkins.

History

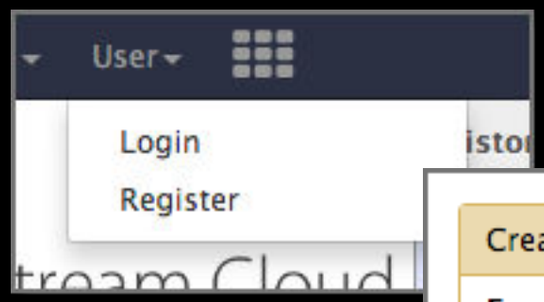
search datasets

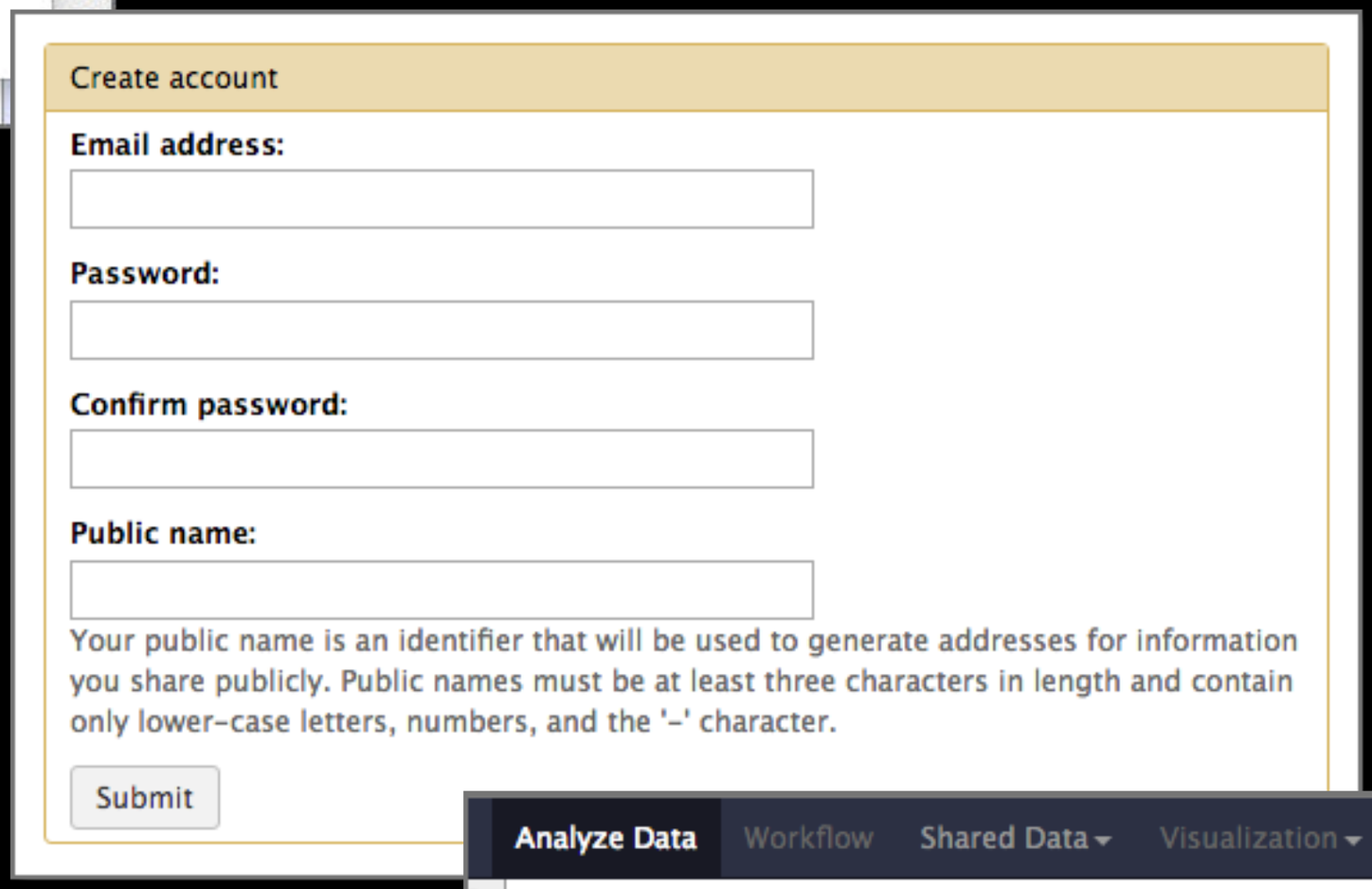
Unnamed history

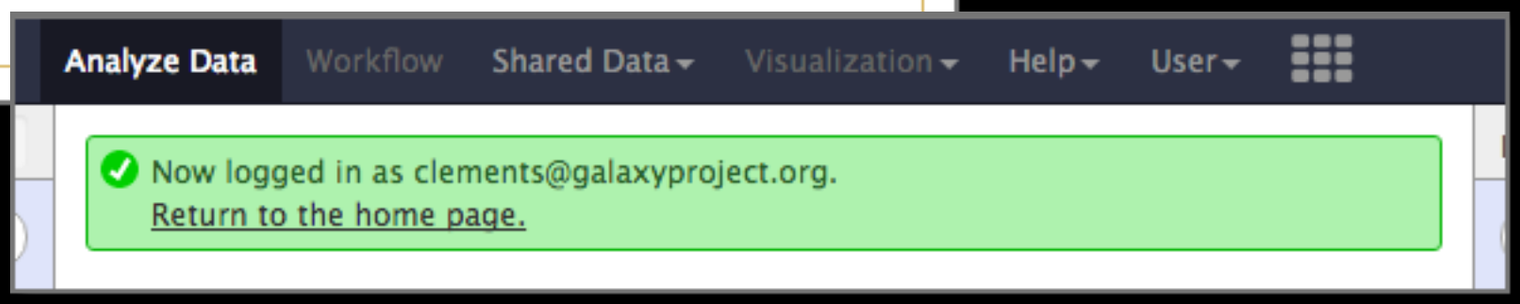
0 b

This history is empty. You can load your own data or get data from an external source

Create a login on your server







Click Return to the home page.

Note: Connection is **not** encrypted

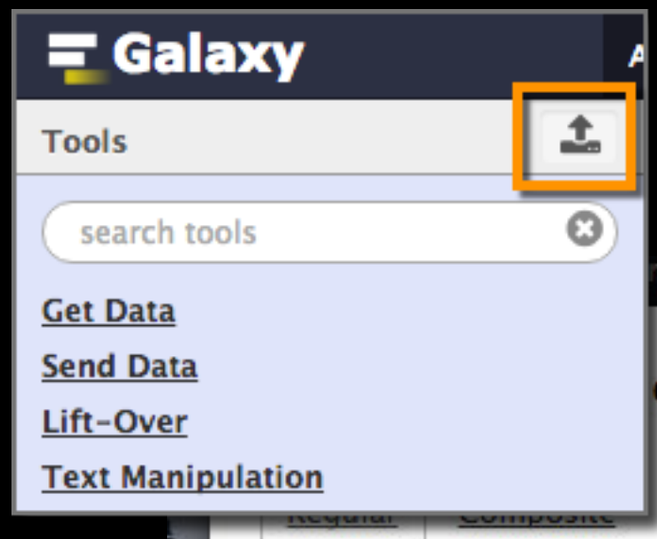
Repeats in Genes: A General Plan

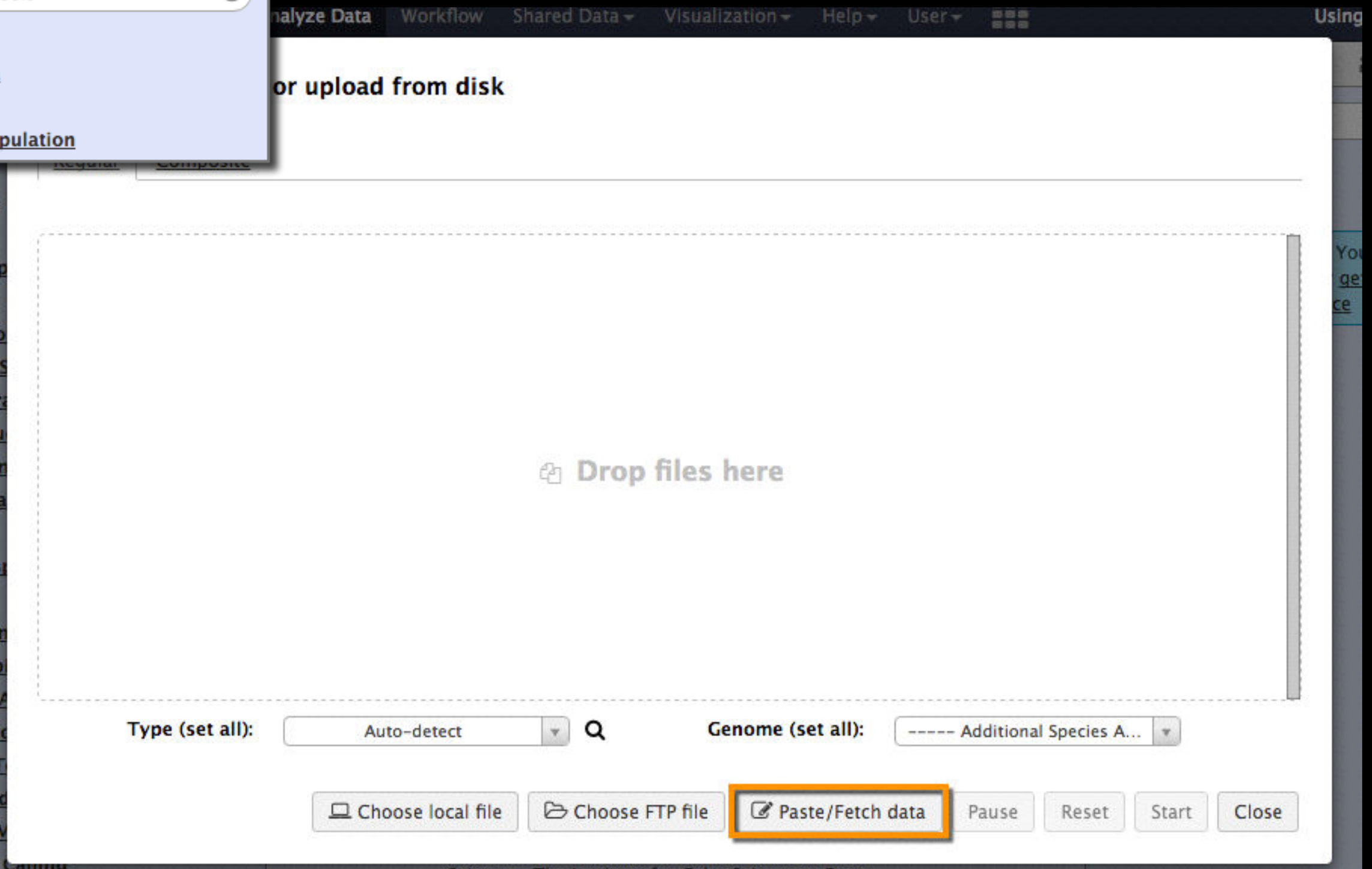
- Get data about Genes, and about Repeats
- Identify which genes have overlapping repeats
- Count repeats per gene

bit.ly/ww_gxy_slides

Get Data: Genes and Repeats

Import 3 data files into your Galaxy instance

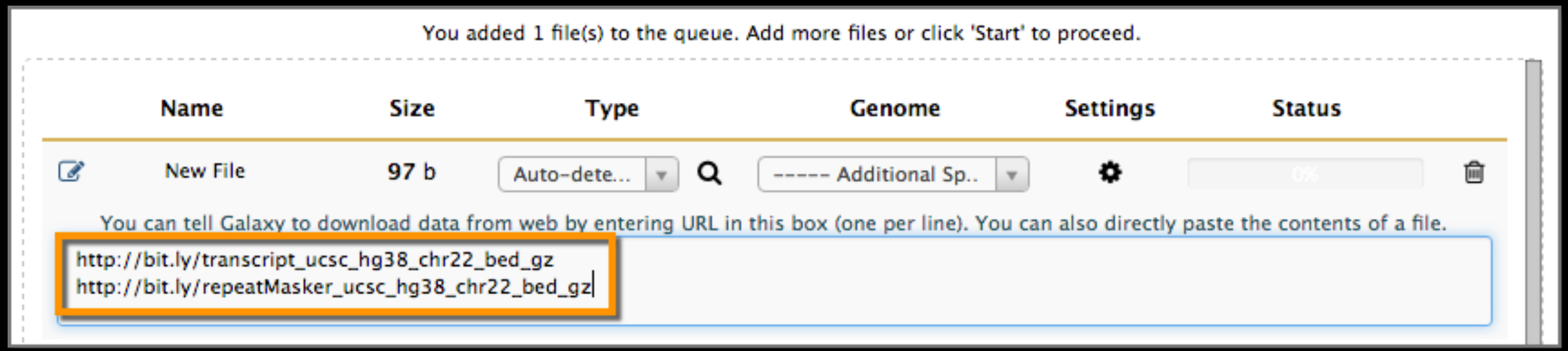




Get Data: Genes and Repeats

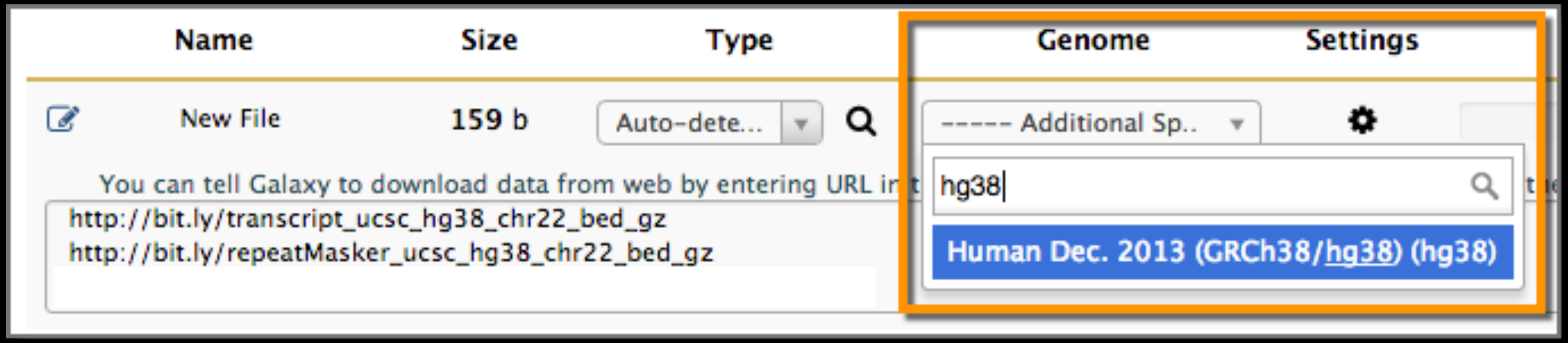
Paste these 2 URLs into the paste box:

http://bit.ly/transcript_ucsc_hg38_chr22_bed_gz
http://bit.ly/repeatMasker_ucsc_hg38_chr22_bed_gz

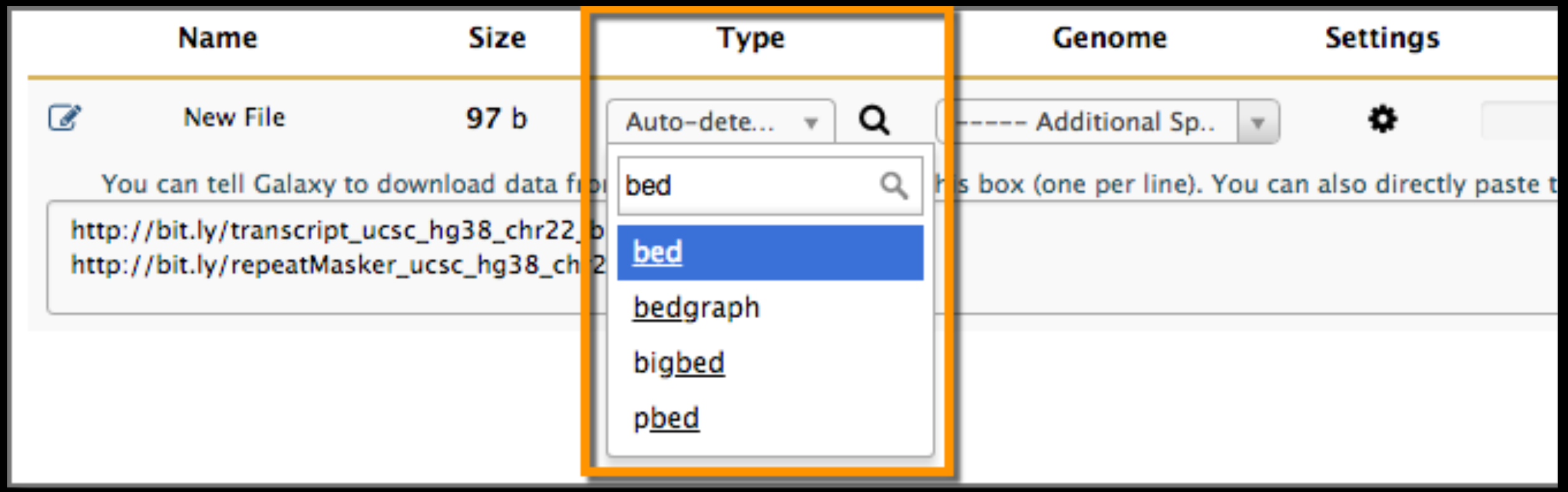


Get Data: Genes and Repeats

Set the Genome to hg38 (the most recent human)



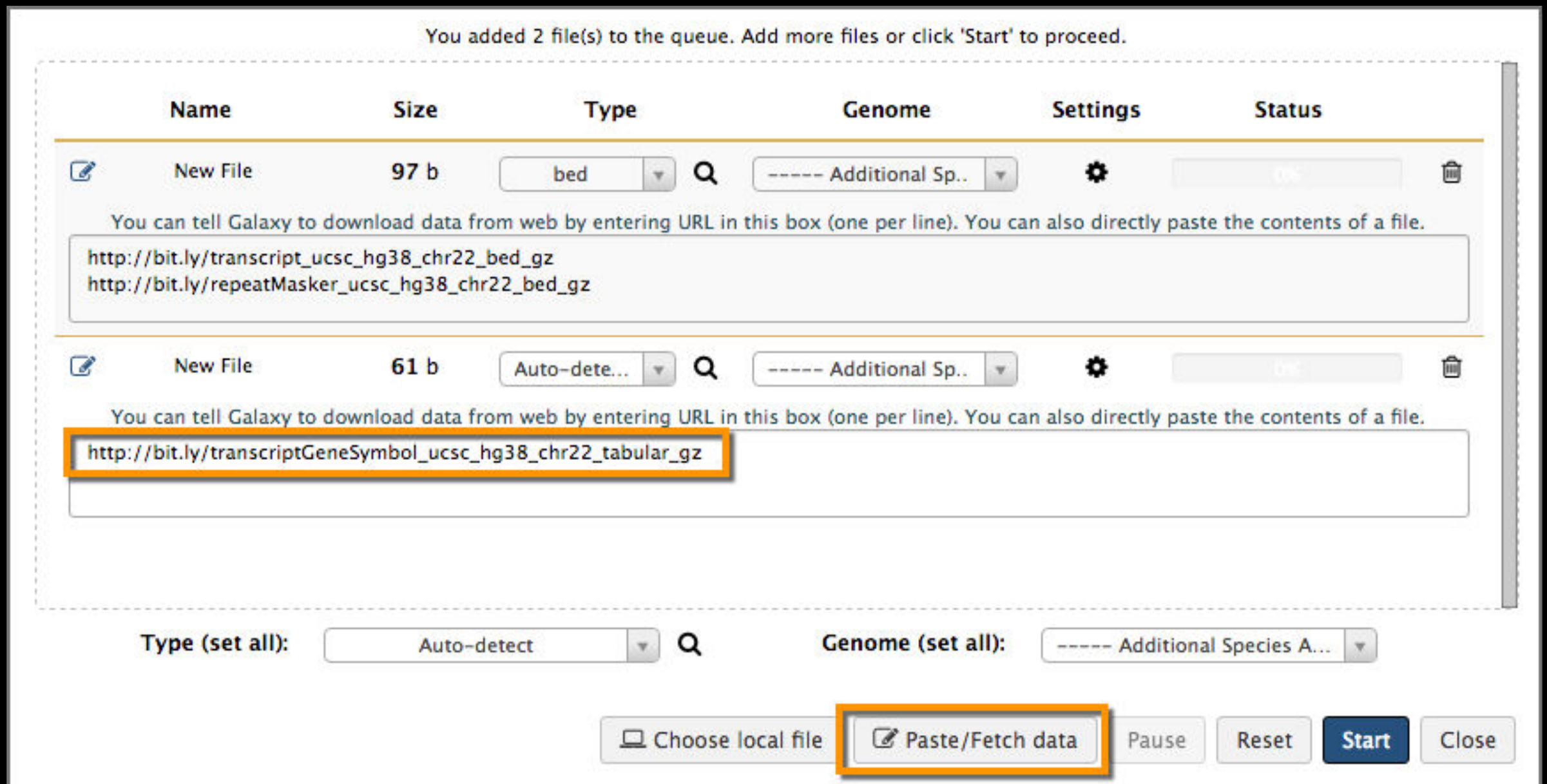
Set the data type to bed



Get Data: Genes and Repeats

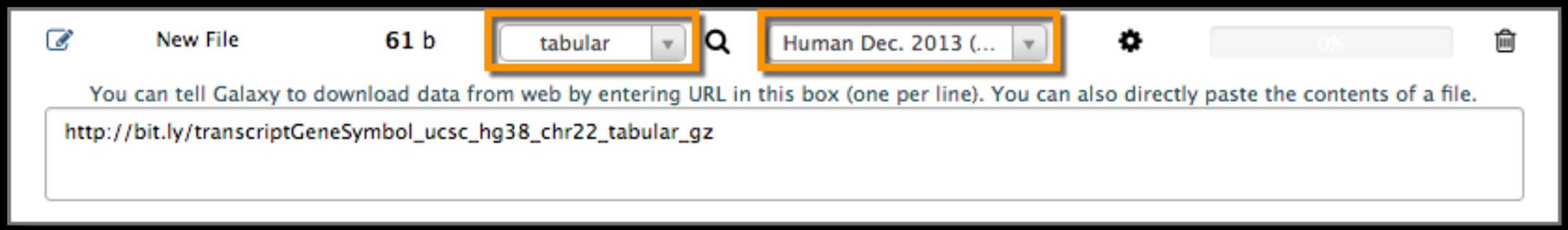
Paste one more dataset. Click Paste/Fetch data again, and then paste this URL into the new paste box:

http://bit.ly/transcriptGeneSymbol_ucsc_hg38_chr22_tabular_gz

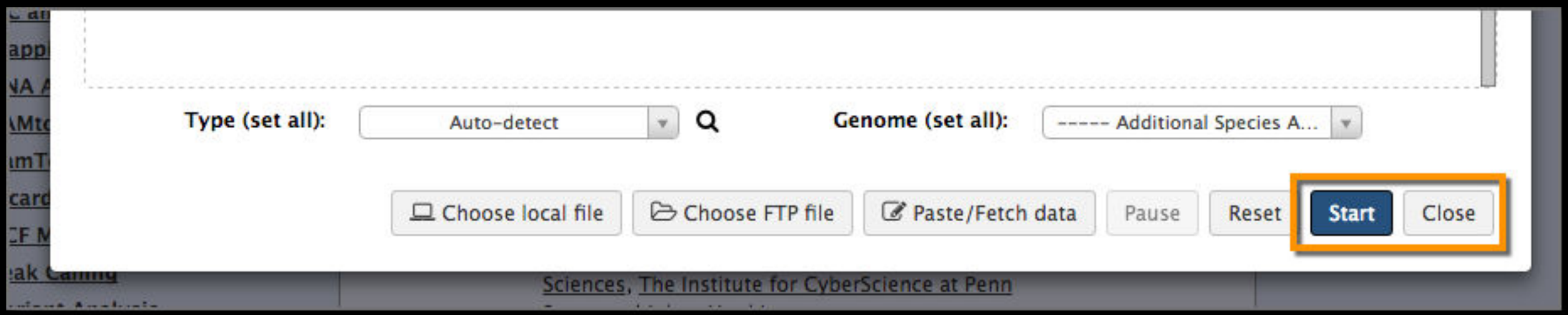


Get Data: Genes and Repeats

Set the data type to **tabular** and the Genome to hg38

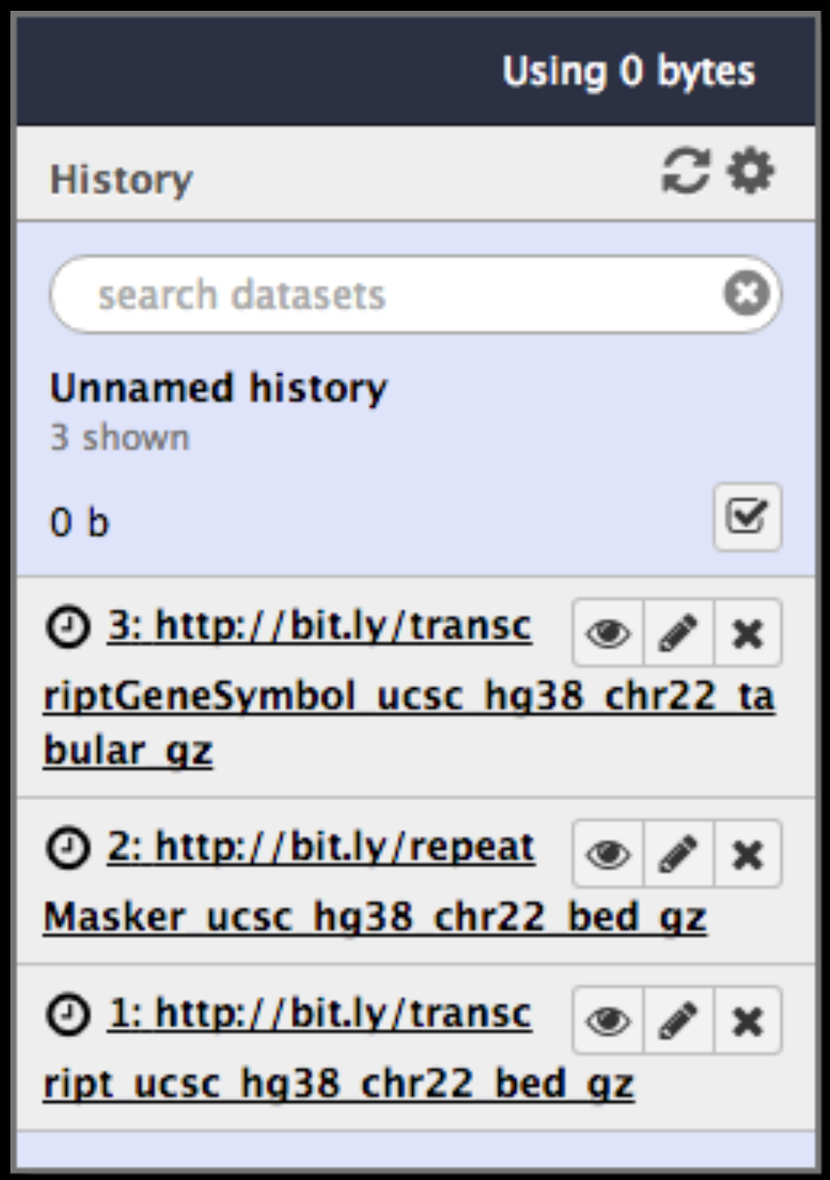


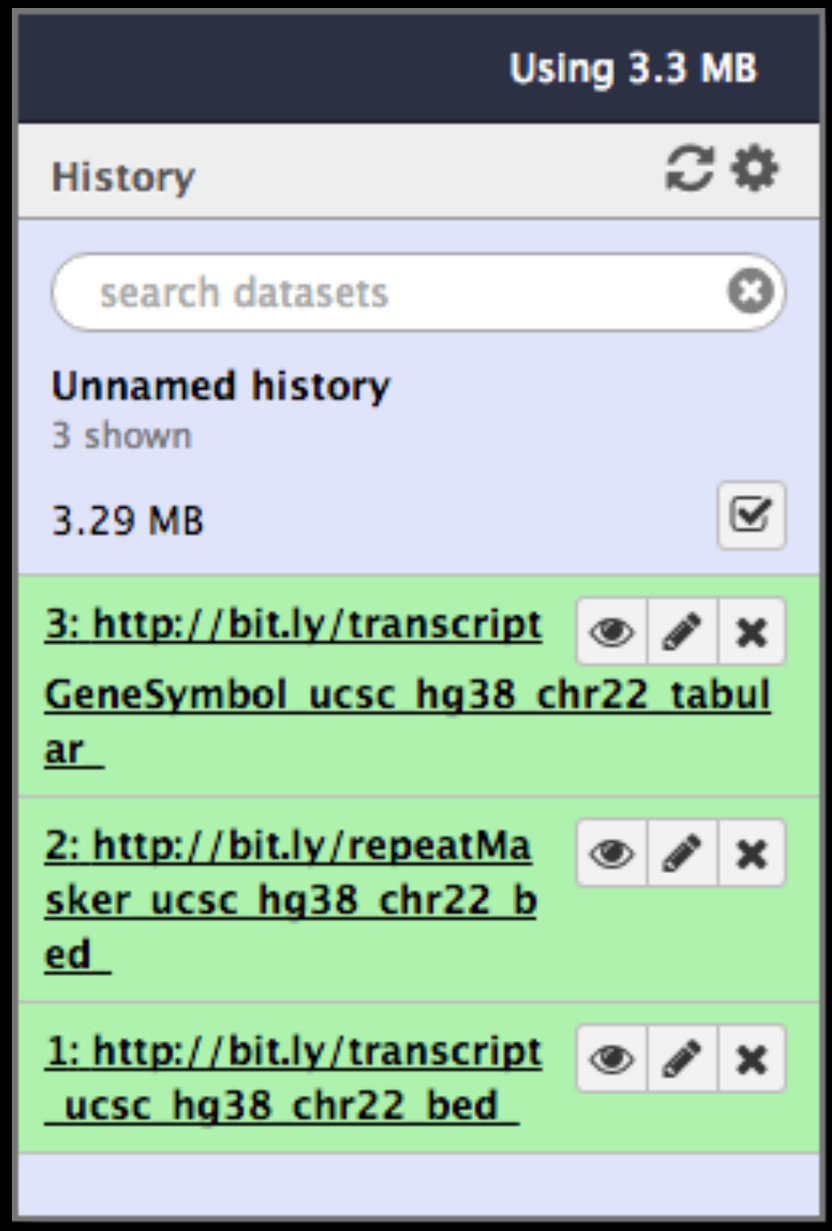
Then click Start, and then Close.

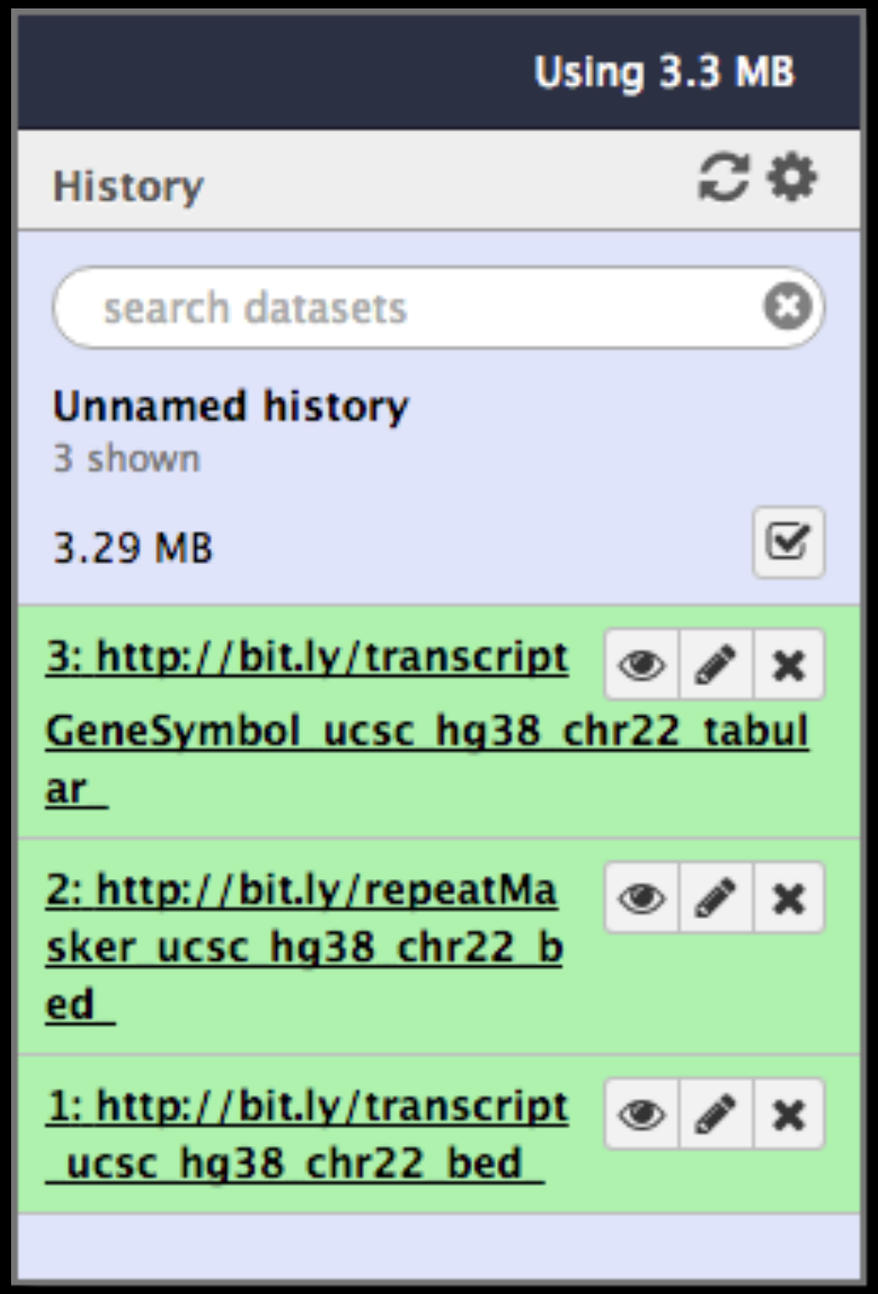


Get Data: Genes and Repeats

The three datasets show up in your history, first as queued, and then as done. The datasets are automatically uncompressed by Galaxy.







Get Data: Datasets

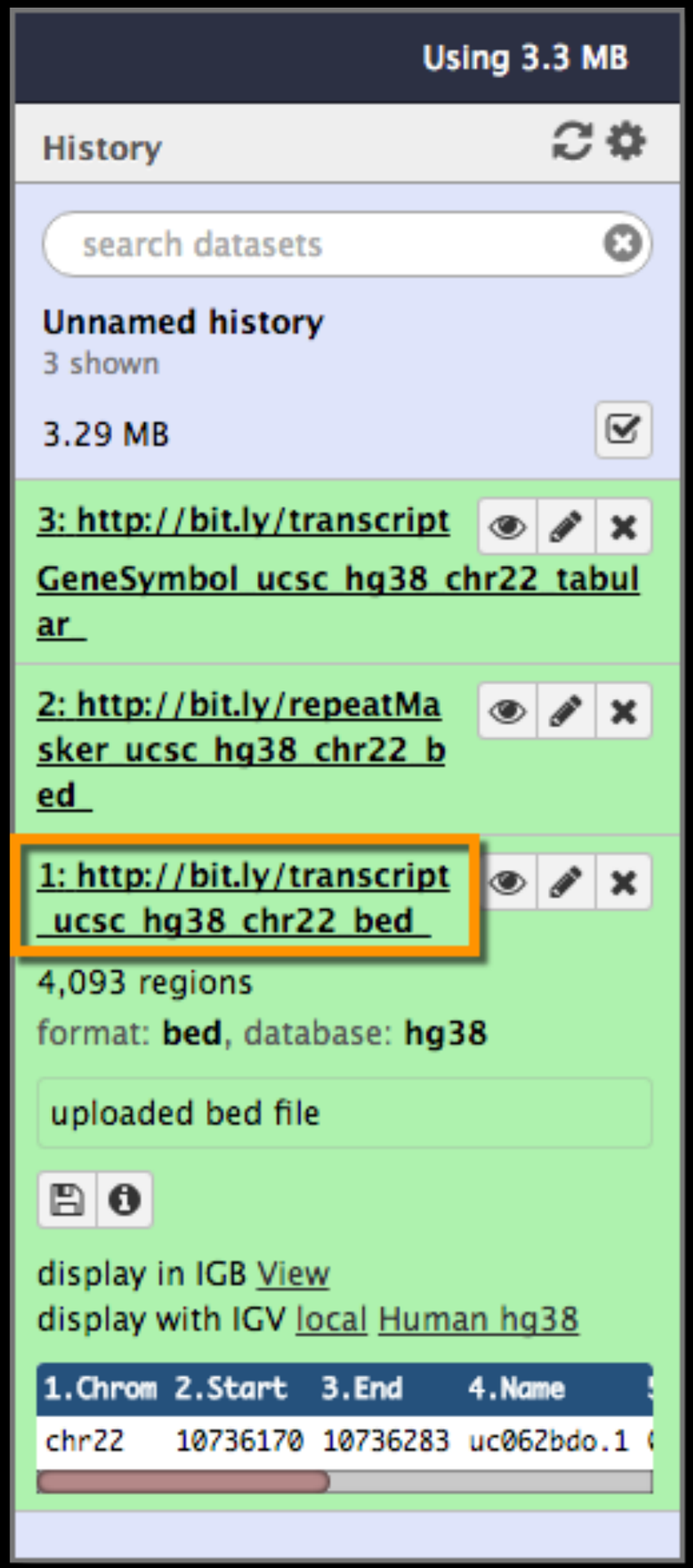
The three datasets are:

3. Transcript to Gene mapping

2. Repeats - as identified by RepeatMasker software

1. Transcripts - Gene and Exon info

Datasets: Take a peek



Preview a dataset in the history by clicking on the dataset's name.

Tells us

How big the dataset is

The format (BED)

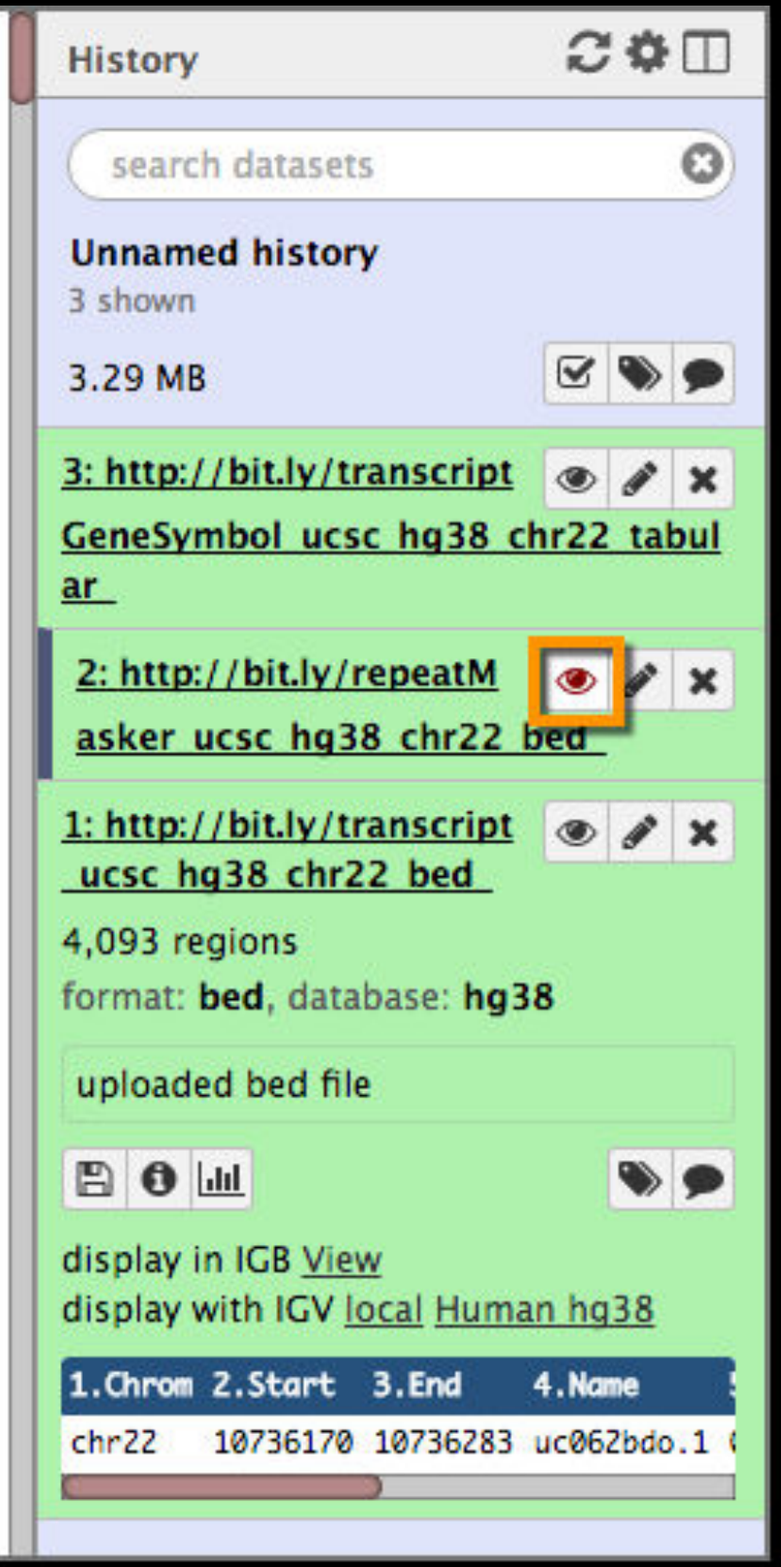
The genome (hg38)

Where it came from

And a short preview of the data in it.

Datasets: See the whole thing (poke it in the eye)

1	2	3	4	5	6
chr22	10510227	10510528	AluSx1	2021	+
chr22	10511018	10511332	L1MC5a	781	-
chr22	10511479	10511791	L1MB1	524	+
chr22	10511878	10512212	L1MB1	313	+
chr22	10512454	10512692	L1MB1	656	+
chr22	10512706	10514778	L1MB1	11092	+
chr22	10514778	10515050	AluSx1	1933	+
chr22	10515050	10515074	L1MB1	11092	+
chr22	10515074	10515121	(GAAG)n	52	+
chr22	10515121	10516103	L1MB1	11092	+
chr22	10516114	10516222	(TA)n	47	+
chr22	10516223	10516285	LTR66	237	-
chr22	10516287	10516630	L1MB1	1504	+
chr22	10516635	10517247	L2a	1062	-
chr22	10517290	10517437	L1MEh	237	-
chr22	10518783	10519114	MLT1A0	1234	+
chr22	10519673	10519746	AluJo	474	-
chr22	10519746	10519816	MER52A	291	-
chr22	10519799	10520945	MER52A	3779	-
chr22	10520950	10521193	AluJo	1258	-
chr22	10522243	10522328	MLT1A0	1424	+
chr22	10522328	10522608	AluSg	2274	+
chr22	10522608	10522644	(AATA)n	39	+
chr22	10522644	10522926	MLT1A0	1424	+



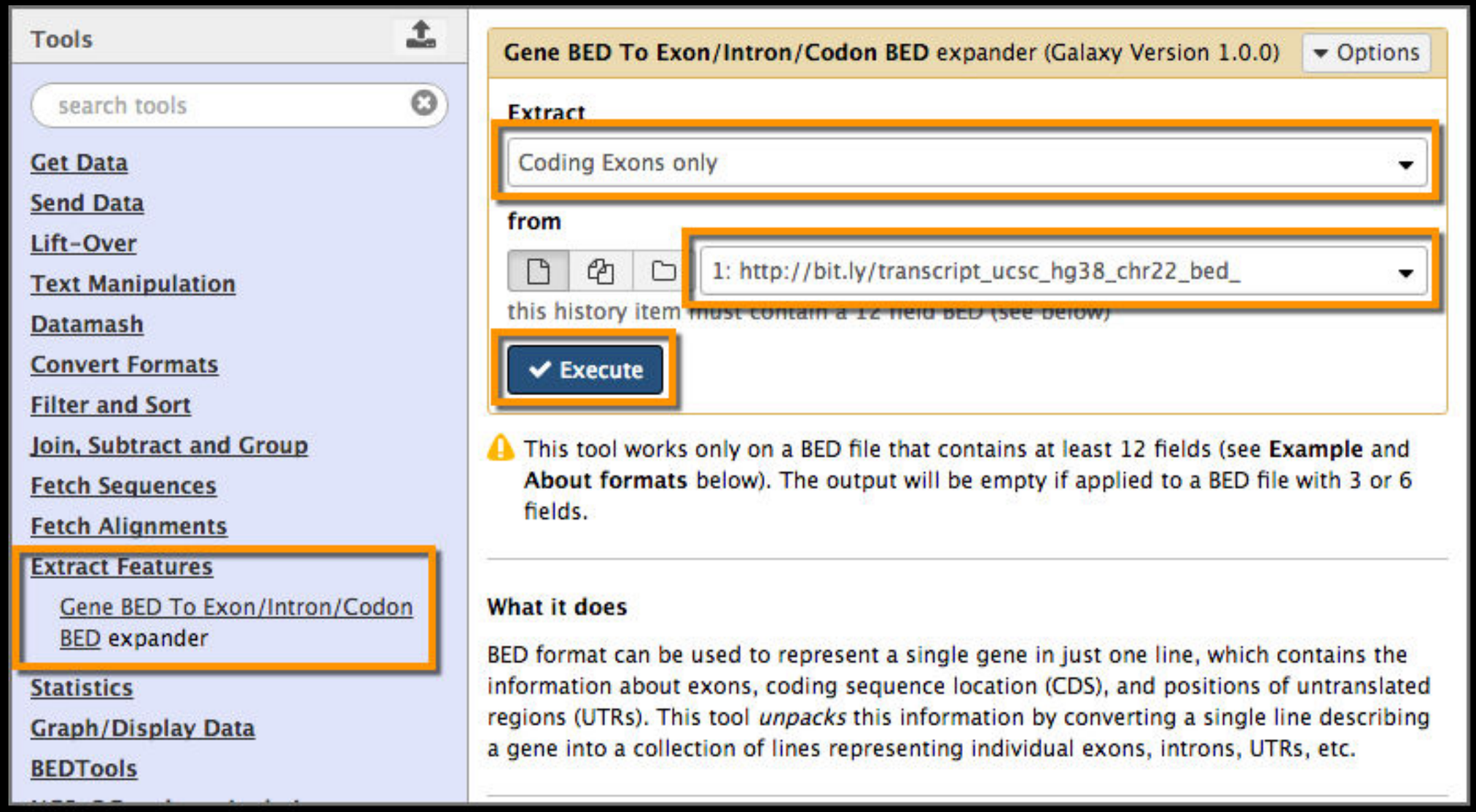
The 6 column RepeatMasker dataset.

Repeats in Genes: A General Plan

- Get data about Genes, and about Repeats
- Identify which genes have overlapping repeats
- Count repeats per gene

bit.ly/ww_gxy_slides

Our first tool: Extract exons from genes

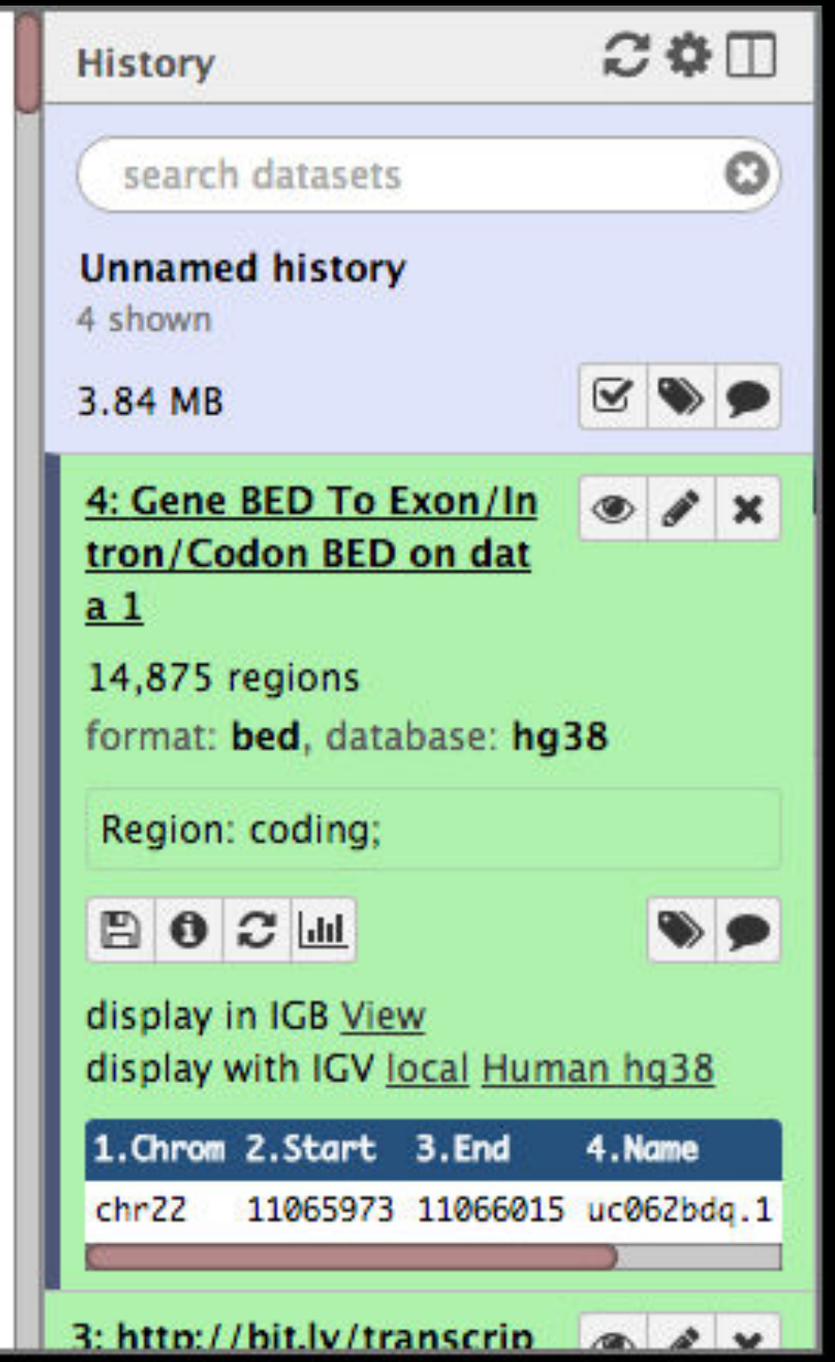


Open the Extract Features toolbox in the tool panel and select Gene BED to Exon/...

Extract Coding Exons only from the transcript dataset. Click Execute.

Extracted Exons

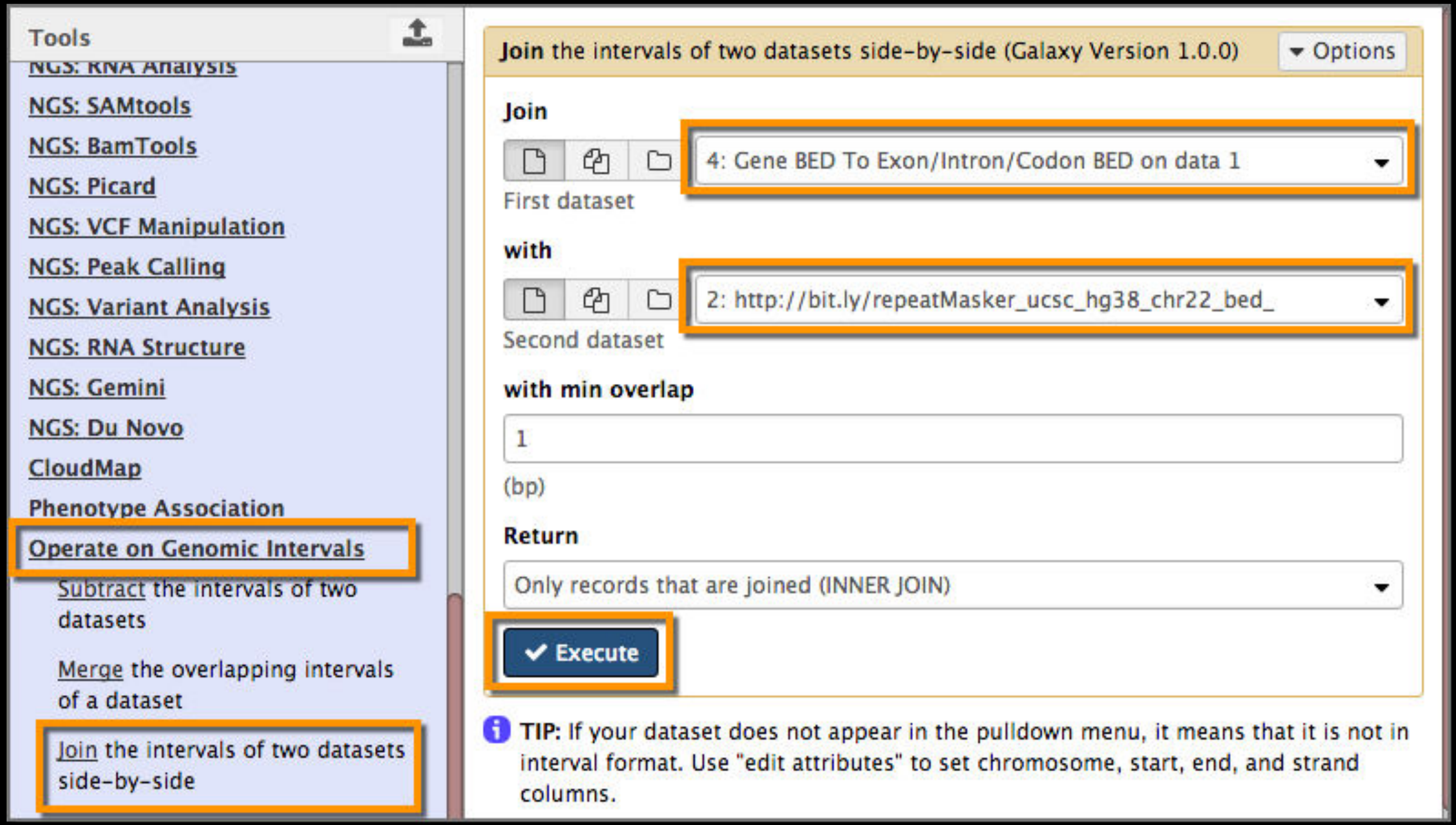
1	2	3	4	5	6
chr22	11065973	11066015	uc062bdq.1	0	-
chr22	11067334	11067346	uc062bdq.1	0	-
chr22	11066500	11066515	uc062bdr.1	0	+
chr22	11067984	11068089	uc062bdr.1	0	+
chr22	15528158	15529139	uc011agd.3	0	+
chr22	15690077	15690709	uc010gqp.3	0	+
chr22	15695370	15695485	uc010gqp.3	0	+
chr22	15695644	15695818	uc010gqp.3	0	+
chr22	15698661	15698768	uc010gqp.3	0	+
chr22	15700077	15700215	uc010gqp.3	0	+
chr22	15702685	15702756	uc010gqp.3	0	+
chr22	15708019	15708090	uc010gqp.3	0	+
chr22	15709781	15709826	uc010gqp.3	0	+
chr22	15710867	15711034	uc010gqp.3	0	+
chr22	15719659	15719777	uc010gqp.3	0	+
chr22	15690077	15690314	uc062bej.1	0	+
chr22	15690425	15690709	uc062bej.1	0	+
chr22	15695370	15695485	uc062bej.1	0	+
chr22	15695644	15695818	uc062bej.1	0	+



6 column dataset, 1 record per exon

Transcript name has become the exon name

Identify which exons have overlapping repeats



Open the Operate on Genomic Intervals toolbox and select Join.
Join the exons dataset with the repeatMasker dataset. Click Execute.

Result has 911 Exon-RepeatMasker pairings

1	2	3	4	5	6	7	8	9	10	11	12
chr22	11065973	11066015	uc062bdq.1	0	-	chr22	11064567	11067436	REP522	6038	+
chr22	11067334	11067346	uc062bdq.1	0	-	chr22	11064567	11067436	REP522	6038	+
chr22	11066500	11066515	uc062bdr.1	0	+	chr22	11064567	11067436	REP522	6038	+
chr22	11067984	11068089	uc062bdr.1	0	+	chr22	11067982	11068155	REP522	348	-
chr22	15697373	15697532	uc062bek.1	0	+	chr22	15697355	15697497	AluJb	1340	-
chr22	15697373	15697532	uc062bek.1	0	+	chr22	15697497	15697530	(TTTTTA)n	32	+
chr22	15697373	15697532	uc062bek.1	0	+	chr22	15697530	15697651	AluJb	1340	-
chr22	17105852	17105954	uc002zly.5	0	+	chr22	17105952	17106169	MIRb	424	+
chr22	17108306	17109820	uc002zly.5	0	+	chr22	17109651	17109678	(GGA)n	15	+
chr22	17108306	17109820	uc062bfr.1	0	+	chr22	17109651	17109678	(GGA)n	15	+
chr22	17119390	17121127	uc002zmb.3	0	-	chr22	17120596	17120630	(GCA)n	15	+
chr22	17119390	17121127	uc002zmb.3	0	-	chr22	17120320	17120347	(GTG)n	15	+
chr22	17119390	17121127	uc002zmb.3	0	-	chr22	17120739	17120793	GA-rich	22	+
chr22	17119390	17121127	uc002zmb.3	0	-	chr22	17120913	17120929	(TGCTGC)n	15	+
chr22	17119390	17121127	uc002zmb.3	0	-	chr22	17120929	17120970	(GCC)n	19	+
chr22	17119390	17121127	uc002zmb.3	0	-	chr22	17120970	17120993	(TGCTGC)n	15	+
chr22	17534743	17534747	uc062bgi.1	0	+	chr22	17534499	17534802	AluY	2037	-
chr22	17550242	17550333	uc062bgm.1	0	+	chr22	17550132	17550323	AluSx1	1189	-
chr22	17619033	17619127	uc062bgu.1	0	-	chr22	17619011	17619306	AluSc8	2210	+
chr22	17726676	17726690	uc010gqy.4	0	+	chr22	17726672	17726726	AmnSINE1	184	-
chr22	17726676	17727534	uc002zmz.4	0	+	chr22	17726672	17726726	AmnSINE1	184	-

Left 6 columns are Exon info, right 6 columns are RepeatMasker info.
Each record represents an overlap pairing.

Repeats in Genes: A General Plan

- Get data about Genes, and about Repeats
- Identify which genes have overlapping repeats
- Count repeats per gene

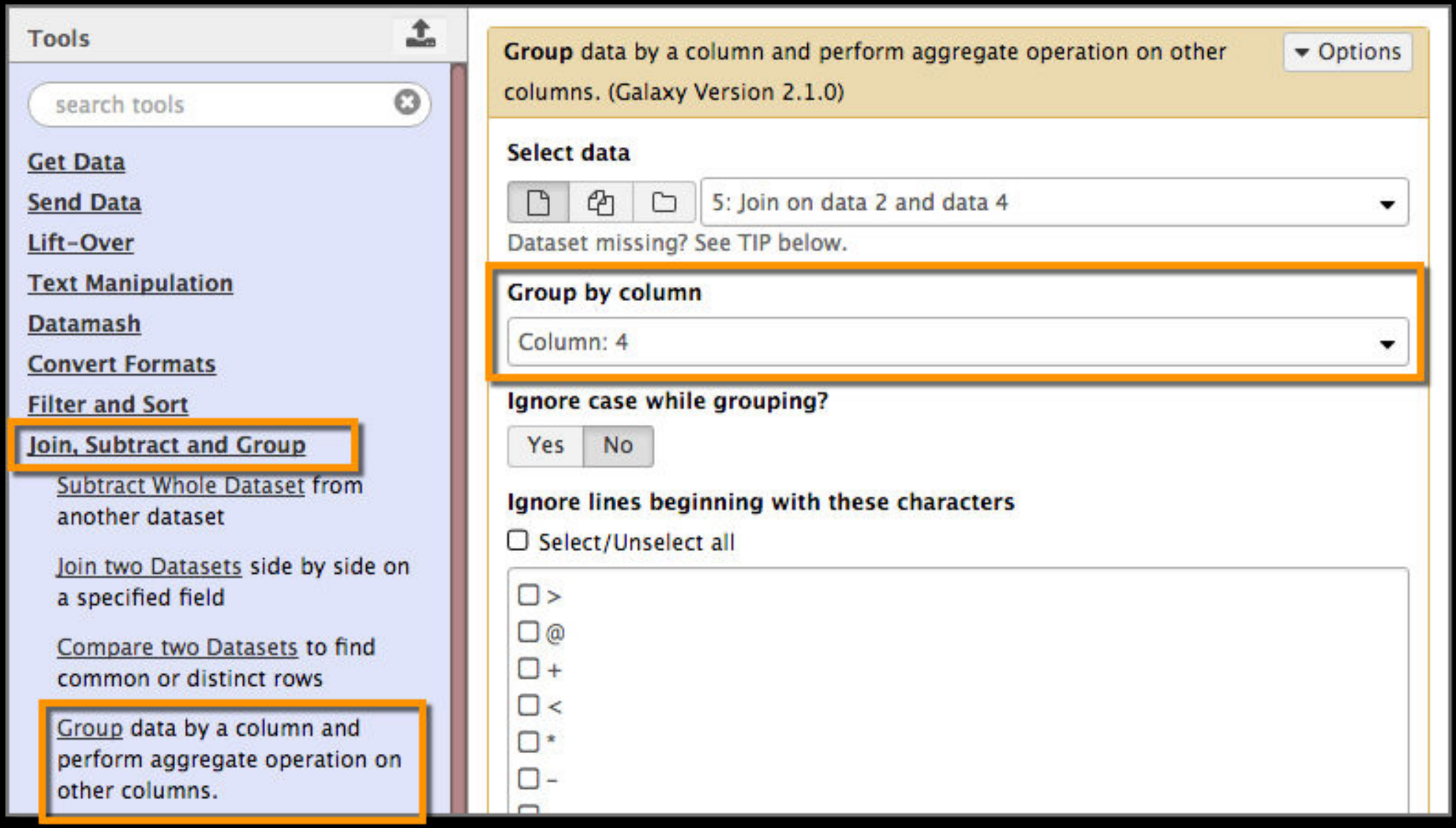
bit.ly/ww_gxy_slides

Group results by Transcript name to get counts

1	2	3	4	5	6	7	8	9	10	11	12
chr22	11065973	11066015	uc062bdq.1	0	-	chr22	11064567	11067436	REP522	6038	+
chr22	11067334	11067346	uc062bdq.1	0	-	chr22	11064567	11067436	REP522	6038	+
chr22	11066500	11066515	uc062bdr.1	0	+	chr22	11064567	11067436	REP522	6038	+
chr22	11067984	11068089	uc062bdr.1	0	+	chr22	11067982	11068155	REP522	348	-
chr22	15697373	15697532	uc062bek.1	0	+	chr22	15697355	15697497	AluJb	1340	-
chr22	15697373	15697532	uc062bek.1	0	+	chr22	15697497	15697530	(TTTTTA)n	32	+
chr22	15697373	15697532	uc062bek.1	0	+	chr22	15697530	15697651	AluJb	1340	-
chr22	17105852	17105954	uc002zly.5	0	+	chr22	17105952	17106169	MIRb	424	+
chr22	17108306	17109820	uc002zly.5	0	+	chr22	17109651	17109678	(GGA)n	15	+
chr22	17108306	17109820	uc062bfr.1	0	+	chr22	17109651	17109678	(GGA)n	15	+
chr22	17119390	17121127	uc002zmb.3	0	-	chr22	17120596	17120630	(GCA)n	15	+
chr22	17119390	17121127	uc002zmb.3	0	-	chr22	17120320	17120347	(GTG)n	15	+
chr22	17119390	17121127	uc002zmb.3	0	-	chr22	17120739	17120793	GA-rich	22	+
chr22	17119390	17121127	uc002zmb.3	0	-	chr22	17120913	17120929	(TGCTGC)n	15	+
chr22	17119390	17121127	uc002zmb.3	0	-	chr22	17120929	17120970	(GCC)n	19	+
chr22	17119390	17121127	uc002zmb.3	0	-	chr22	17120970	17120993	(TGCTGC)n	15	+
chr22	17534743	17534747	uc062bgi.1	0	+	chr22	17534499	17534802	AluY	2037	-
chr22	17550242	17550333	uc062bgm.1	0	+	chr22	17550132	17550323	AluSx1	1189	-
chr22	17619033	17619127	uc062bgu.1	0	-	chr22	17619011	17619306	AluSc8	2210	+
chr22	17726676	17726690	uc010gqy.4	0	+	chr22	17726672	17726726	AmnSINE1	184	-
chr22	17726676	17727534	uc002zmz.4	0	+	chr22	17726672	17726726	AmnSINE1	184	-

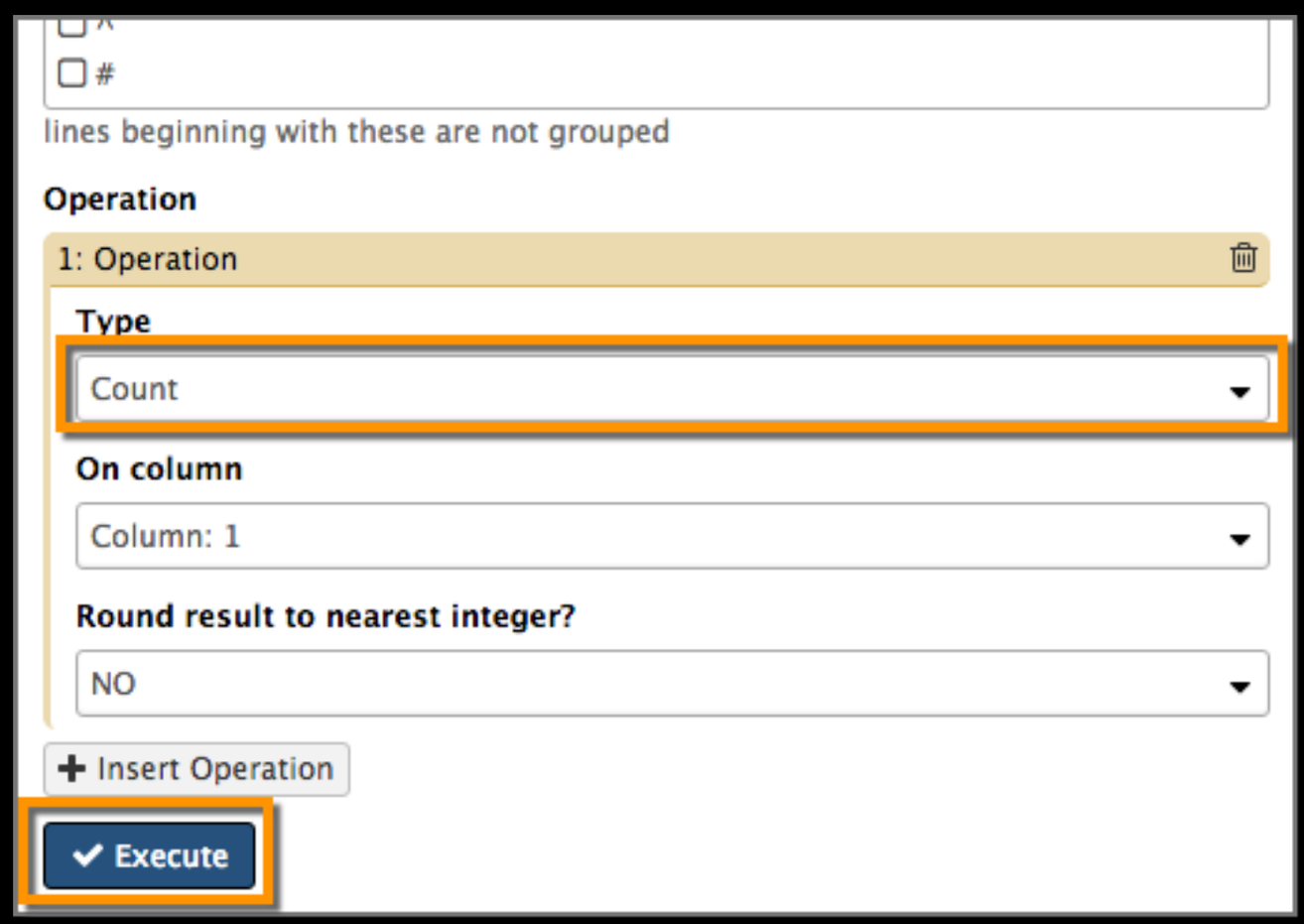
There is a record for every time a repeat overlaps an exon in a transcript.
The # of records with each transcript name is the # of overlaps.

Group results by Transcript name to get counts



Open Join, Subtract and Group toolbox and select Group. Set Group by column to Column 4 (the transcript name).

Group results by Transcript name to get counts



Click + Insert Operation and set the Type to Count. Click Execute.
For each value of column 4, keep a count of # records with that value.

Group results by Transcript name returns:

1	2
uc002zly.5	2
uc002zmb.3	6
uc002zmw.5	1
uc002zmx.4	1
uc002zmy.5	1
uc002zmz.4	1
uc002zng.5	3
uc002zny.4	1
uc002zou.4	1
uc002zoy.5	1
uc002zoz.6	2
uc002zpf.2	1
uc002zpi.4	2
uc002zpk.2	1
uc002zpm.3	1
uc002zpo.3	1
uc002zqa.2	4
uc002zqb.4	2
uc002zqc.4	2

Returns 628 transcripts that have one or more overlapping repeats.

But, biologists more often think in terms of Genes, rather than transcripts.

Let's associate these counts with the transcript's genes.

3rd imported dataset is a mapping from transcript names to genes.

1	2
#hg38.knownGene.name	hg38.kgXref.geneSymbol
uc062bdo.1	U2
uc062bdp.1	CU459211.1
uc062bdq.1	CU104787.1
uc062bdr.1	BAGE5
uc062bds.1	5_8S_rRNA
uc062bdt.1	AC137488.1
uc062bdu.1	AC137488.2
uc062bdv.1	CU013544.1
uc062bdw.1	CT867976.1
uc062bdx.1	CT867977.1
uc062bdy.1	CT978678.1
uc062bdz.1	CU459202.1
uc062bea.1	AC116618.1
uc062beb.1	CU463998.1
uc062bec.1	CU463998.3
uc062bed.1	CU463998.2
uc062bee.1	U6
uc062bef.1	LA16c-60D12.1
uc062beg.1	LA16c-60D12.2
uc011agd.3	OR11H1
uc062beh.1	LA16c-23H5.4
uc062bei.1	LA16c-2F2.8
uc010gqp.3	POTEH
uc062bej.1	POTEH
uc062bek.1	POTEH
uc002zlk.5	POTEH-AS1
uc062bel.1	RNU6-816P
uc062bem.1	LINC01297

History

search datasets

Unnamed history
6 shown

3.92 MB

6: Group on data 5

5: Join on data 2 and dat a 4

4: Gene BED To Exon/Int ron/Codon BED on data 1

3: http://bit.ly/transcrip tGeneSymbol_ucsc_hg38_chr22_tab ular

4,093 lines, 1 comments
format: **tabular**, database: **hg38**

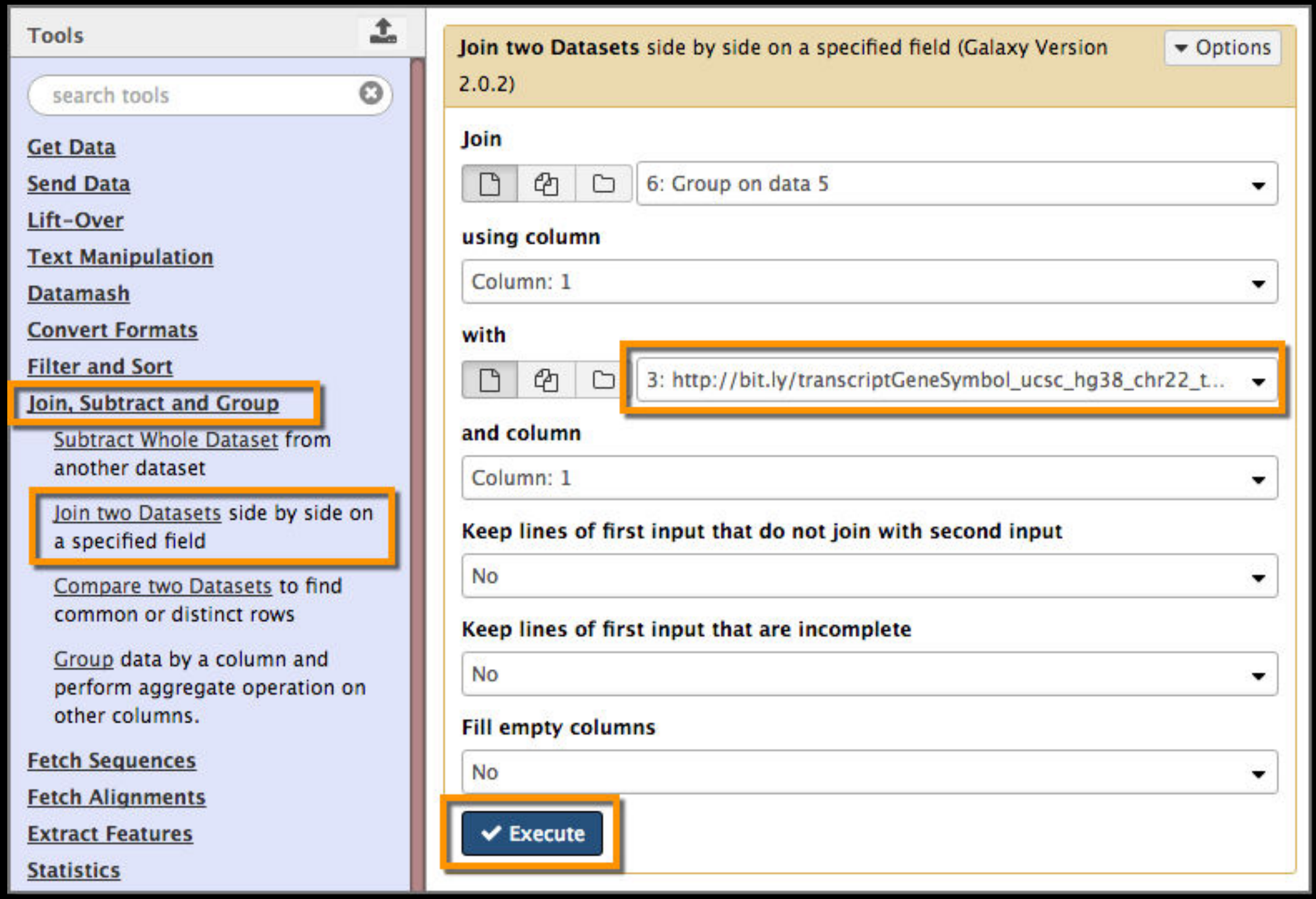
uploaded tabular file

1	2
#hg38.knownGene.name	hg38.kgXref.gene

2: http://bit.ly/repeatMa sker_ucsc_hg38_chr22_b ed

1: http://bit.ly/transcript _ucsc_hg38_chr22_bed

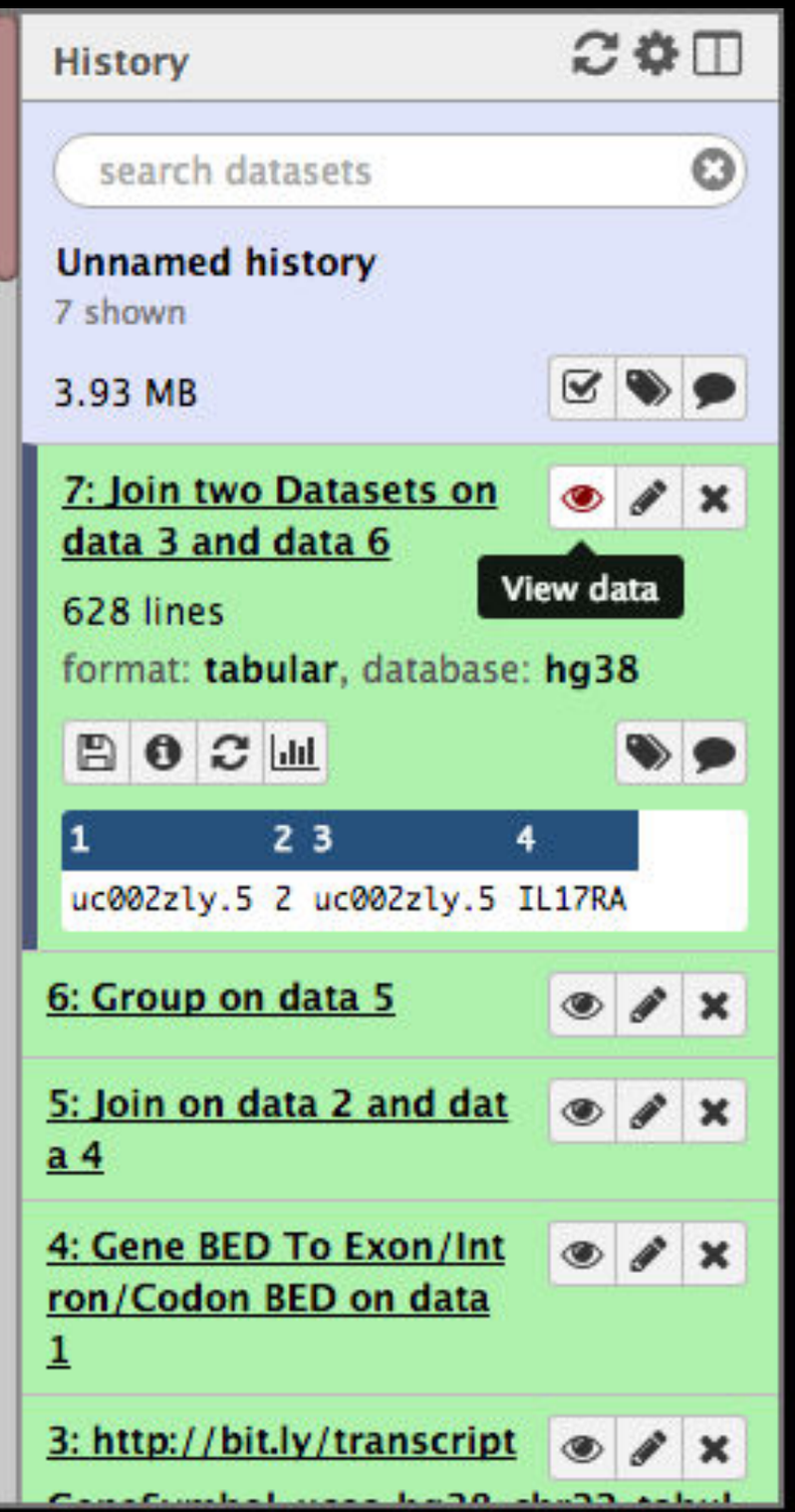
Associate transcript counts with genes



Open Join, Subtract and Group toolbox and select Join. Join the transcript count dataset (Group on) with the dataset 3 (transcriptGeneSymbol mapping file). Transcript name is Column 1 in both datasets.

For each transcript, we now have a count, and gene name

1	2	3	4
uc002zly.5	2	uc002zly.5	IL17RA
uc002zmb.3	6	uc002zmb.3	CECR6
uc002zmw.5	1	uc002zmw.5	BCL2L13
uc002zmx.4	1	uc002zmx.4	BCL2L13
uc002zmy.5	1	uc002zmy.5	BCL2L13
uc002zmz.4	1	uc002zmz.4	BCL2L13
uc002zng.5	3	uc002zng.5	MICAL3
uc002zny.4	1	uc002zny.4	USP18
uc002zou.4	1	uc002zou.4	DGCR14
uc002zoy.5	1	uc002zoy.5	SLC25A1
uc002zoz.6	2	uc002zoz.6	SLC25A1
uc002zpf.2	1	uc002zpf.2	HIRA
uc002zpi.4	2	uc002zpi.4	C22orf39
uc002zpk.2	1	uc002zpk.2	C22orf39
uc002zpm.3	1	uc002zpm.3	UFD1L
uc002zpo.3	1	uc002zpo.3	UFD1L
uc002zqa.2	4	uc002zqa.2	TBX1
uc002zqb.4	2	uc002zqb.4	TBX1
uc002zqc.4	2	uc002zqc.4	TBX1
uc002zqg.4	1	uc002zqg.4	C22orf29
uc002zqh.4	1	uc002zqh.4	C22orf29
uc002zqi.4	1	uc002zqi.4	C22orf29



For some genes, we have different overlap counts.

For each gene, get the max # of overlaps

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help | User

Tools

search tools

- Get Data
- Send Data
- Lift-Over
- Text Manipulation
- Datamash
- Convert Formats
- Filter and Sort
- Join, Subtract and Group
  - Subtract Whole Dataset from another dataset
  - Join two Datasets side by side on a specified field
  - Compare two Datasets to find common or distinct rows
  - Group data by a column and perform aggregate operation on other columns.

Group data by a column and perform aggregate operation on other columns. (Galaxy Version 2.1.0) — Options

Select data

7: Join two Datasets on data 3 and data 6

Dataset missing? See TIP below.

Group by column

Column: 4

Ignore case while grouping?

Yes | No

Ignore lines beginning with these characters

Select/Unselect all

Operation

1: Operation

Type

Maximum

On column

Column: 2

Round result to nearest integer?

NO

+ Insert Operation

Execute

Why column 4 and column 2?

230 Genes on chr22 have 1+ overlapping repeats

8: Group on data 7

230 lines
format: **tabular**, database: **hg38**

--Group by c4: max[c2]

1	2
AC007326.1	4

Sort the results so genes with most overlaps are sorted first.

Break ties by gene name.

Tools

search tools

Get Data
Send Data
Lift-Over
Text Manipulation
Datamash
Convert Formats
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
Filter GTF data by attribute values_list
Join, Subtract and Group
Fetch Sequences

Sort data in ascending or descending order (Galaxy Version 1.0.3) — Options

Sort Dataset: 8: Group on data 7

on column: Column: 2

with flavor: Numerical sort

everything in: Descending order

Column selection

1: Column selection

on column: Column: 1

with flavor: Alphabetical sort

everything in: Ascending order

+ Insert Column selection

✔ Execute

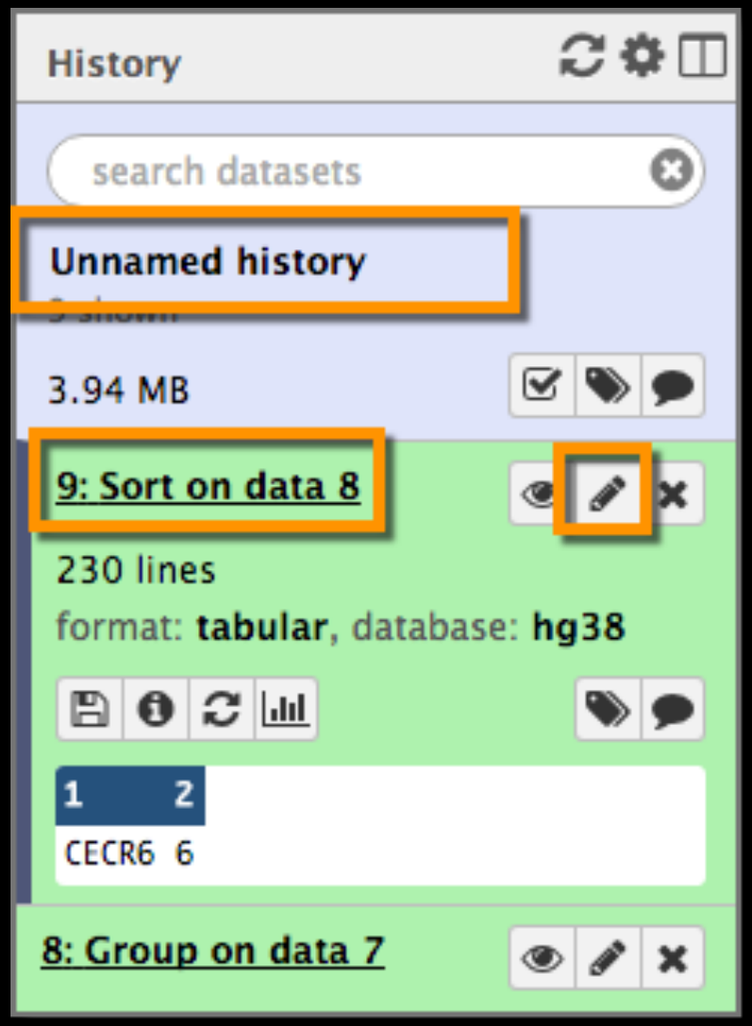
Maximum # overlaps on chr22 is 6 in 3 different genes

1	2
CECR6	6
SCARF2	6
SHANK3	6
MYH9	5
TCF20	5
AC007326.1	4
BAIAP2L2	4
CACNA1I	4
CRELD2	4
DRICH1	4
MAPK8IP2	4
MED15	4
RIMBP3	4
RIMBP3C	4
SH3BP1	4
TBX1	4
TMEM191B	4
TNRC6B	4
CABIN1	3
CARD10	3
CCDC157	3
CELSR1	3

Before we move on, a word about naming

Unnamed history
and
Sort on data 8
are both, um, accurate, but not informative.

A best practice is to name your histories, your inputs, and outputs with informative names.



To name the history, click on Unnamed history, enter the new name, and then hit the Return key.

To rename a dataset click on the dataset's pencil (edit attributes) icon. (And see the next slide.)

Rename a dataset

Attributes | Convert Format | Datatype | Permissions

Edit Attributes

Name:

Genes w # overlapping repeats, chr2

Info:

Sort on data 8

Annotation / Notes:

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

Database/Build:

Human Dec. 2013 (GRCh38/hg38) (hg38)

Number of comment lines:

Save

History

search datasets

Genes-RepeatMasker overlaps, chr22

9 shown

3.94 MB

9: Sort on data 8

230 lines

format: **tabular**, database: **hg38**

1 2

CECR6 6

8: Group on data 7

7: Join two Datasets on data 3 and data 6

6: Group on data 5

5: Join on data 2 and dat

Agenda

Launching Galaxy in Jetstream

A quick stroll through the Galaxy

Demonstrate Galaxy by addressing a specific question

Turn that analysis into a reusable workflow

Shutting down Galaxy in Jetstream

Galaxy Ecosystem

bit.ly/ww_gxy_slides

Now, let's rerun that analysis

With

- entire genome
- different species
- repeats identified by other repeat software
- ...

Reselecting all those tools and parameters manually is error prone.

It's also a path to insanity.

Some Galaxy Terminology

Dataset:

Any input, output or intermediate set of data + metadata

History:

A series of inputs, analysis steps, intermediate datasets, and outputs

Workflow:

A series of analysis steps

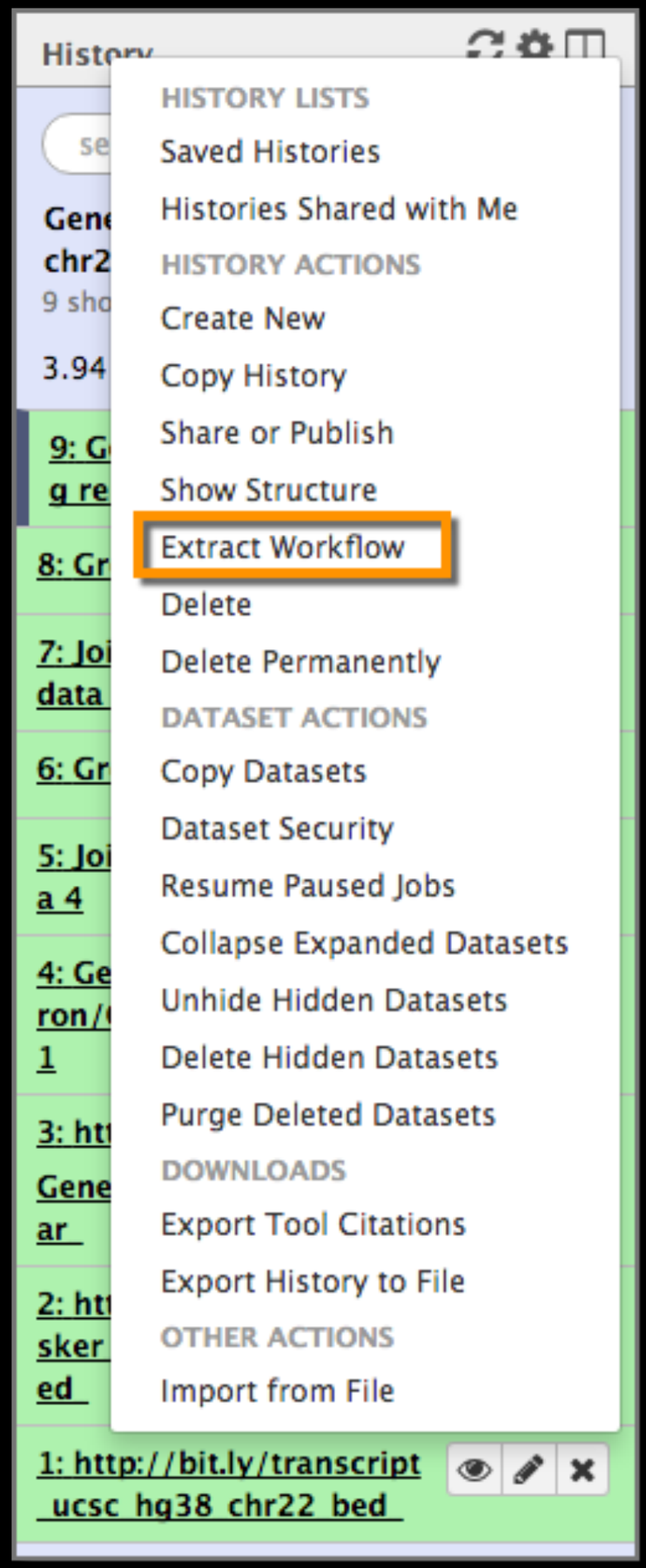
Can be repeated with different data

Create a Workflow from a History

Extract Workflow from history

Create a workflow from this history.
Edit it to make some things clearer.

(cog) → Extract Workflow



Create a Workflow from a History

Give it a meaningful name and click **Create Workflow.**

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

Genes-Repeats overlaps

Create Workflow | Check all | Uncheck all

Tool	History items created
Upload File *This tool cannot be used in workflows*	1: http://bit.ly/transcript_ucsc_hg38_chr22_bed Treat as input dataset
	2: http://bit.ly/repeatMasker_ucsc_hg38_chr22_bed Treat as input dataset
Upload File *This tool cannot be used in workflows*	3: http://bit.ly/transcriptGeneSymbol_ucsc_hg38_chr22_tabular Treat as input dataset
Gene BED To Exon/Intron/Codon BED Include "Gene BED To Exon/Intron/Codon BED" in workflow	4: Gene BED To Exon/Intron/Codon BED on data 1
Join Include "Join" in workflow	5: Join on data 2 and data 4
Group Include "Group" in workflow	6: Group on data 5
Join two Datasets Include "Join two Datasets" in workflow	7: Join two Datasets on data 3 and data 6
Group Include "Group" in workflow	8: Group on data 7
Sort Include "Sort" in workflow	9: Genes w # overlapping repeats, chr22

History

search datasets

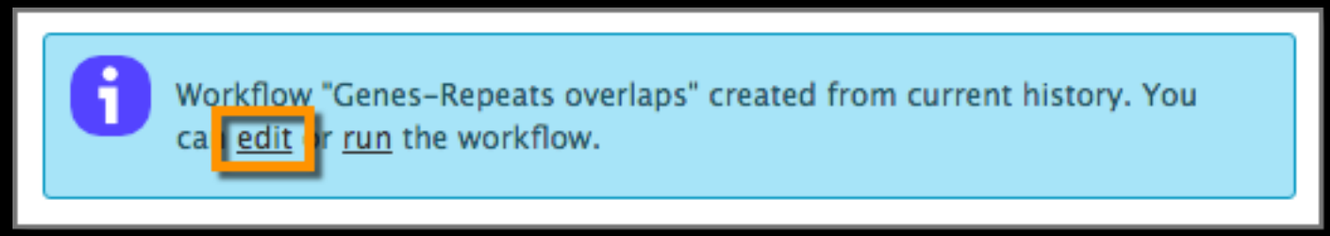
Genes-RepeatMasker overlaps, chr22

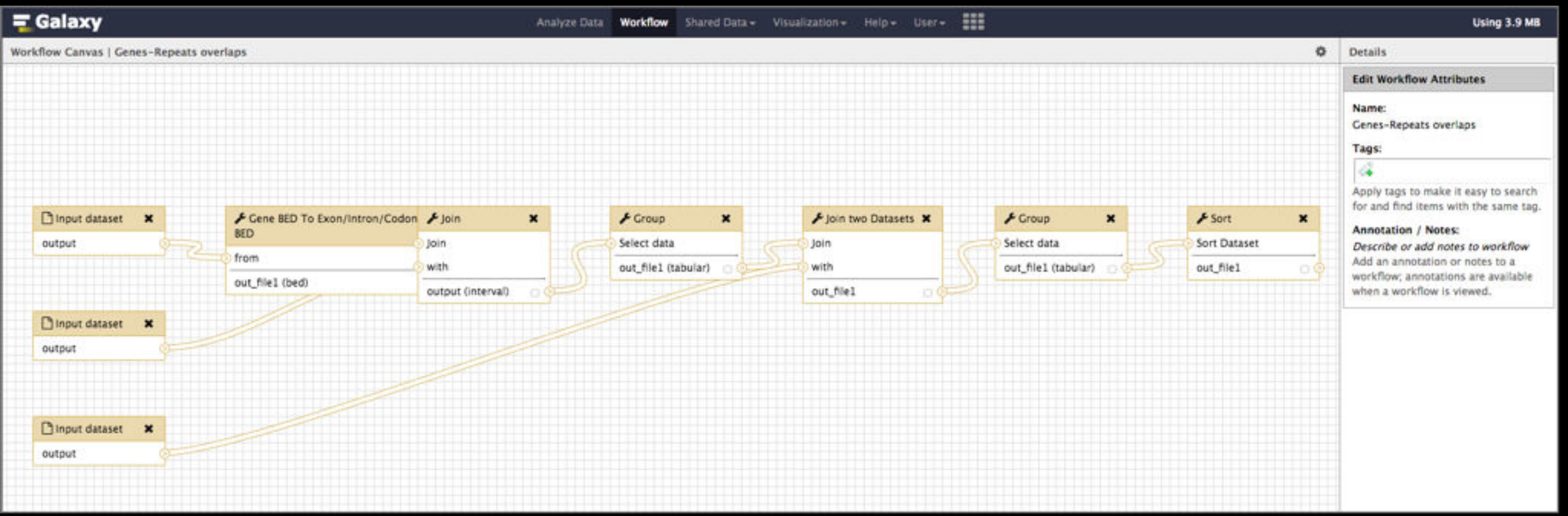
9 shown

3.94 MB

- 9: Genes w # overlapping repeats, chr22
- 8: Group on data 7
- 7: Join two Datasets on data 3 and data 6
- 6: Group on data 5
- 5: Join on data 2 and data 4
- 4: Gene BED To Exon/Intron/Codon BED on data 1
- 3: http://bit.ly/transcriptGeneSymbol_ucsc_hg38_chr22_tabular
- 2: http://bit.ly/repeatMasker_ucsc_hg38_chr22_bed
- 1: http://bit.ly/transcript_ucsc_hg38_chr22_bed

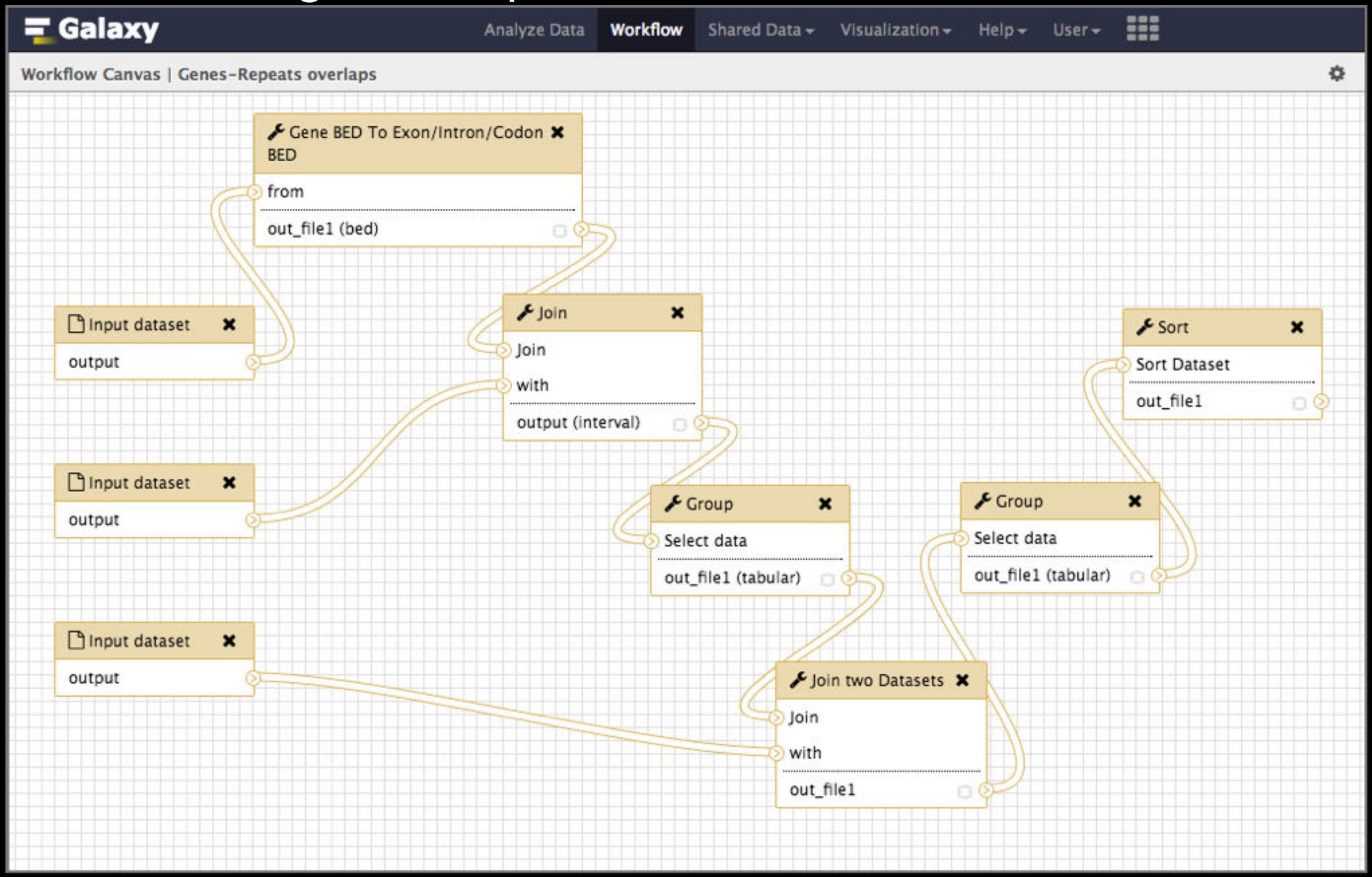
Edit the workflow



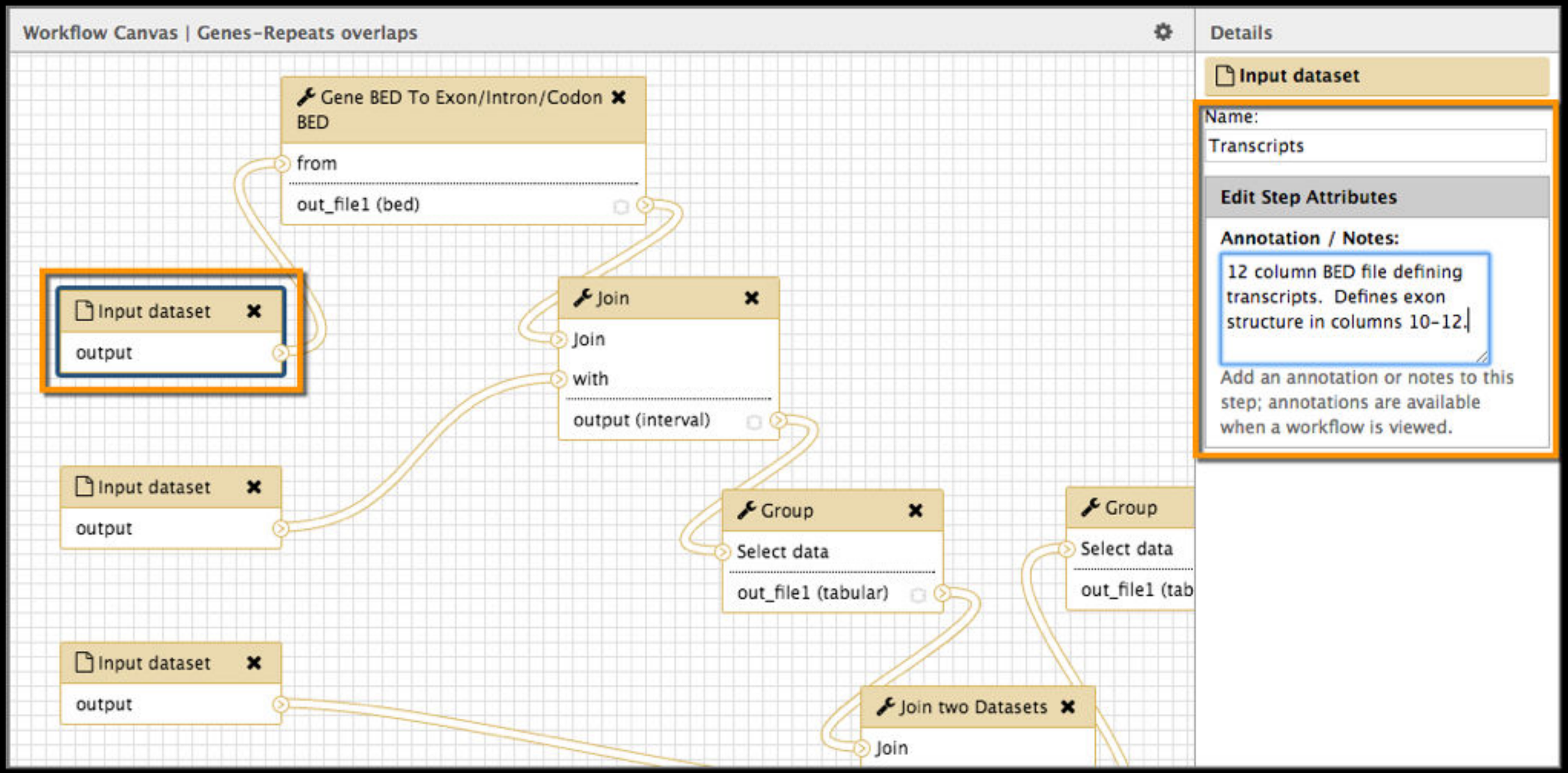


After resizing the window, and collapsing the tool panel, this is the initial layout of the workflow.

Rearrange the steps to make the data flow clearer.

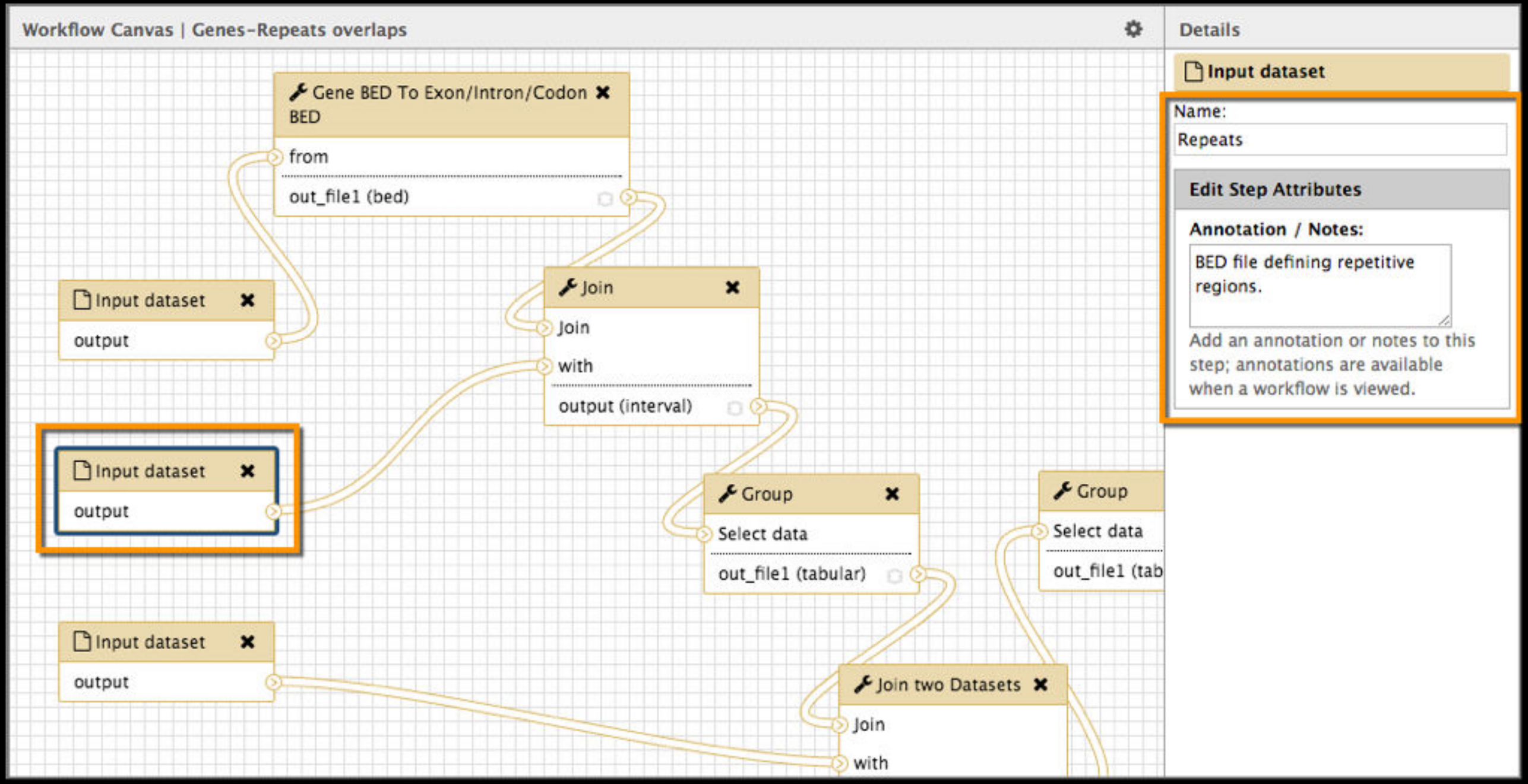


Give the input and output datasets meaningful names



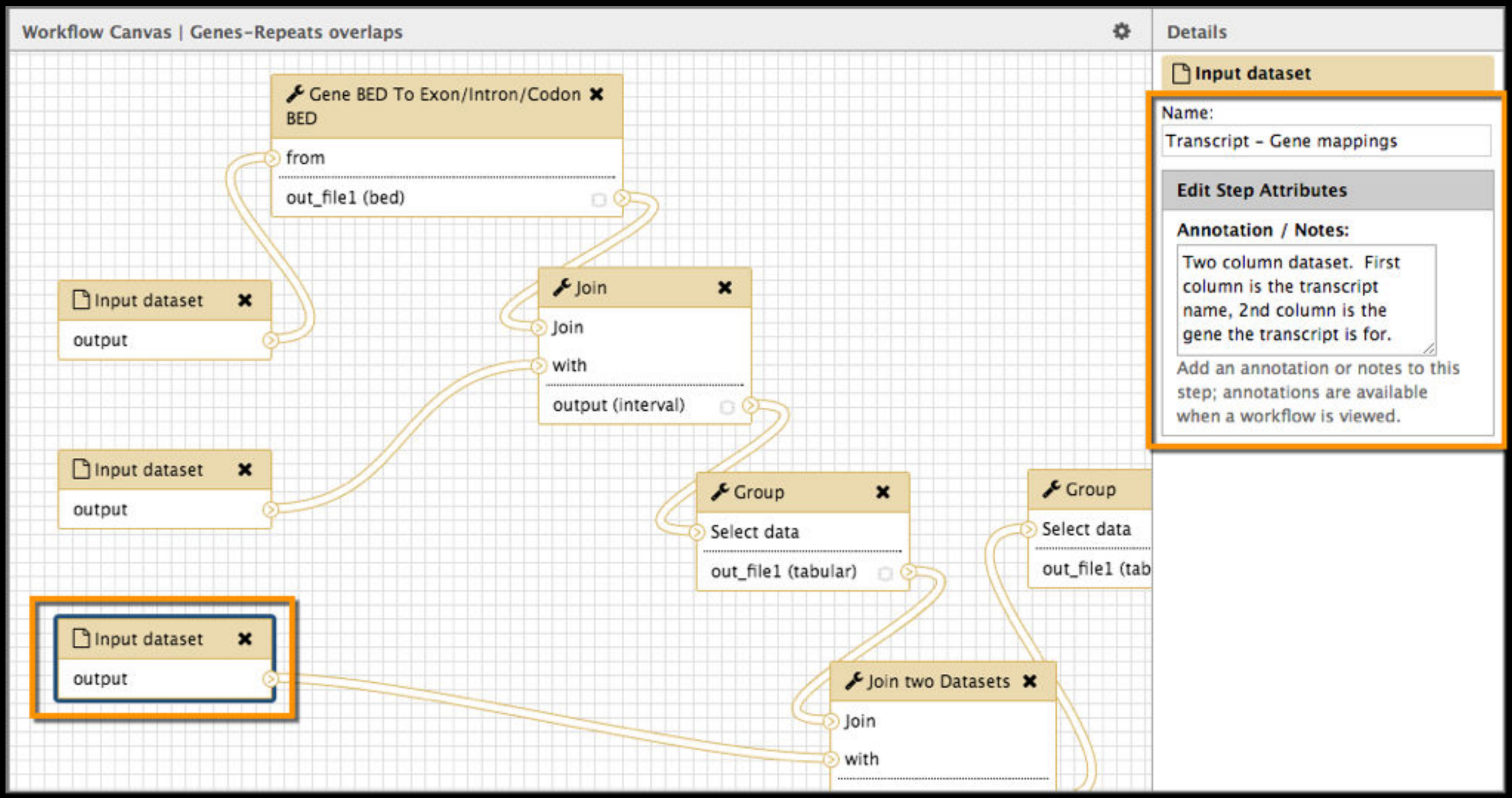
Name the transcripts input dataset.

Give the input and output datasets meaningful names



Name the repeats input dataset.

Give the input and output datasets meaningful names



Name the Transcript Gene mapping input dataset.

Name the output dataset.

Workflow Canvas | Genes-Repeats overlaps

BED To Exon/Intron/Codon

(bed)

Join

Join

with

output (interval)

Group

Select data

out_file1 (tabular)

Group

Select data

out_file1 (tabular)

Join two Datasets

Join

with

out_file1

Sort

Sort Dataset

out_file1

Details

Alphabetical sort

everything in

Ascending order

Insert Column selection

Annotation / Notes

Genes with maximum number of overlapping repeats in any of the gene's transcripts.

step. It will be shown with the workflow.

Email notification

Yes | No

An email notification will be send when the job has completed.

Output cleanup

Yes | No

Delete intermediate outputs if they are not used as input for another job.

Configure Output: 'out_file1'

Label

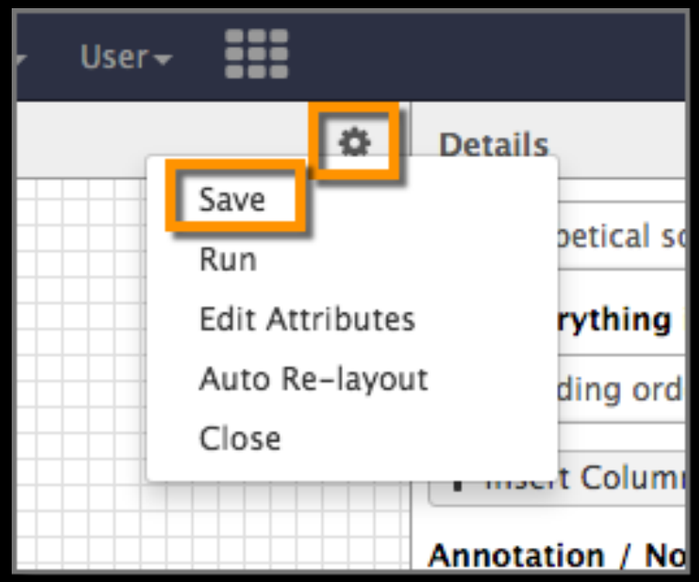
This will provide a short name to describe the output – this must be unique across workflows.

Rename dataset

Genes w max # of overlapping repe

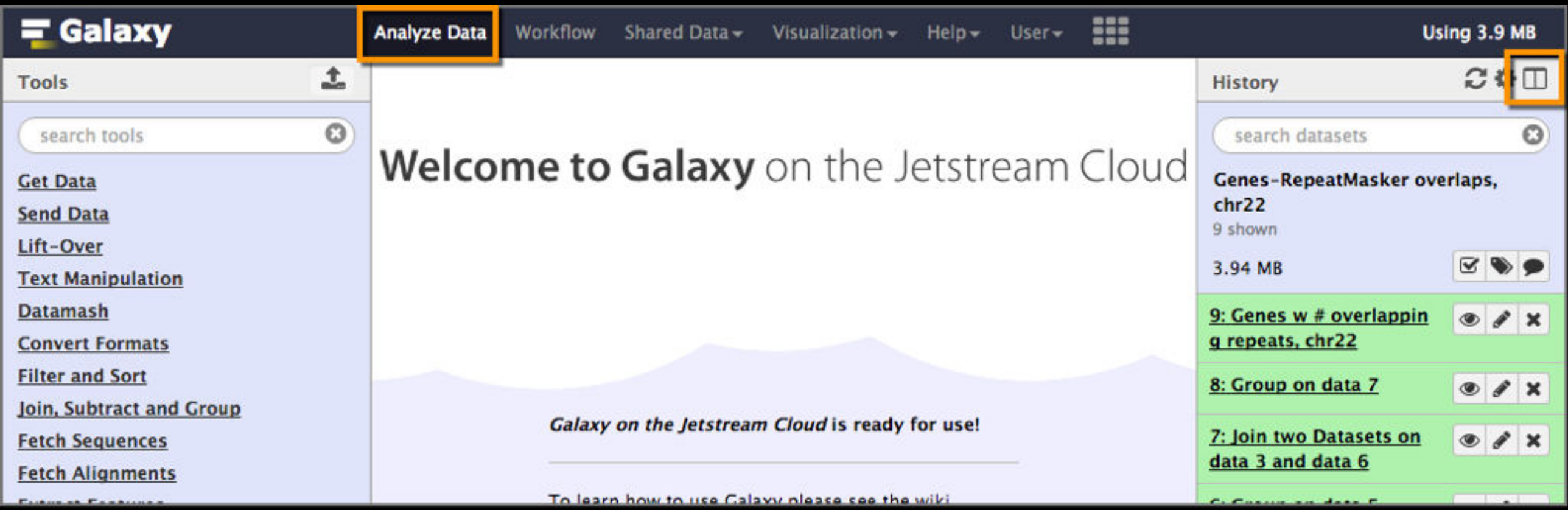
dataset. Click here for more

Save your workflow edits!



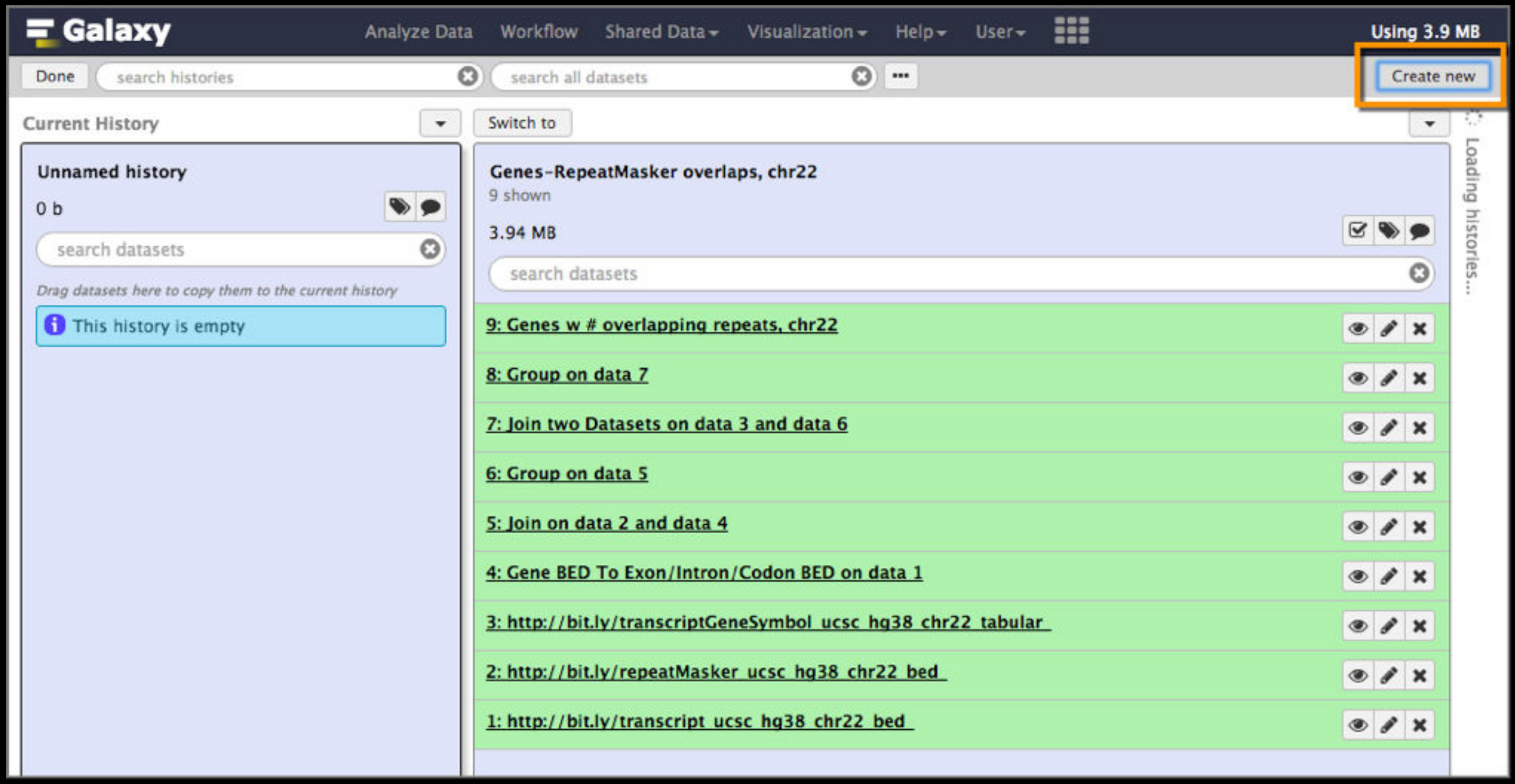
Review your workflows

Start a new history



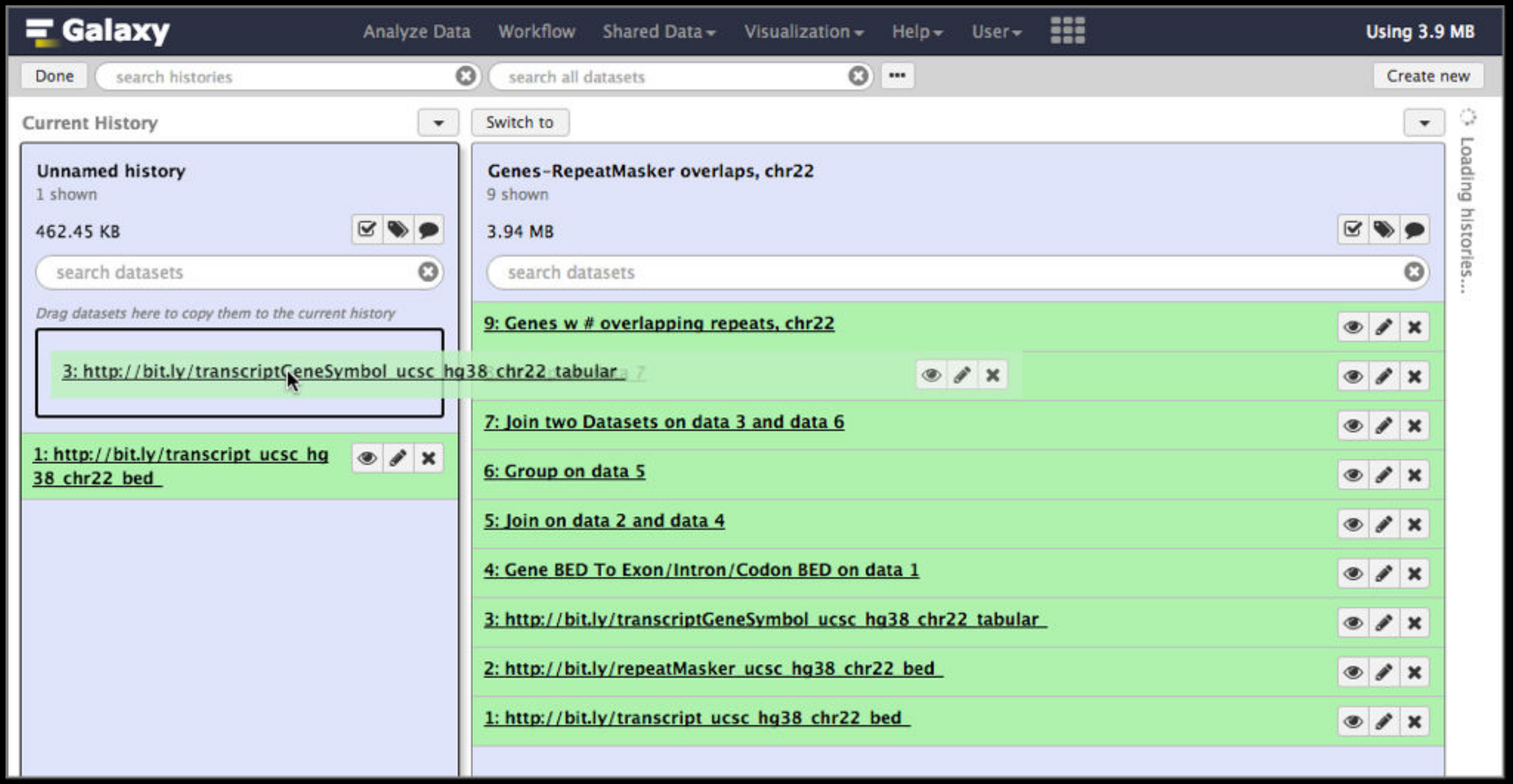
Click the Analyze Data tab and then
click the View all histories icon

Start a new history

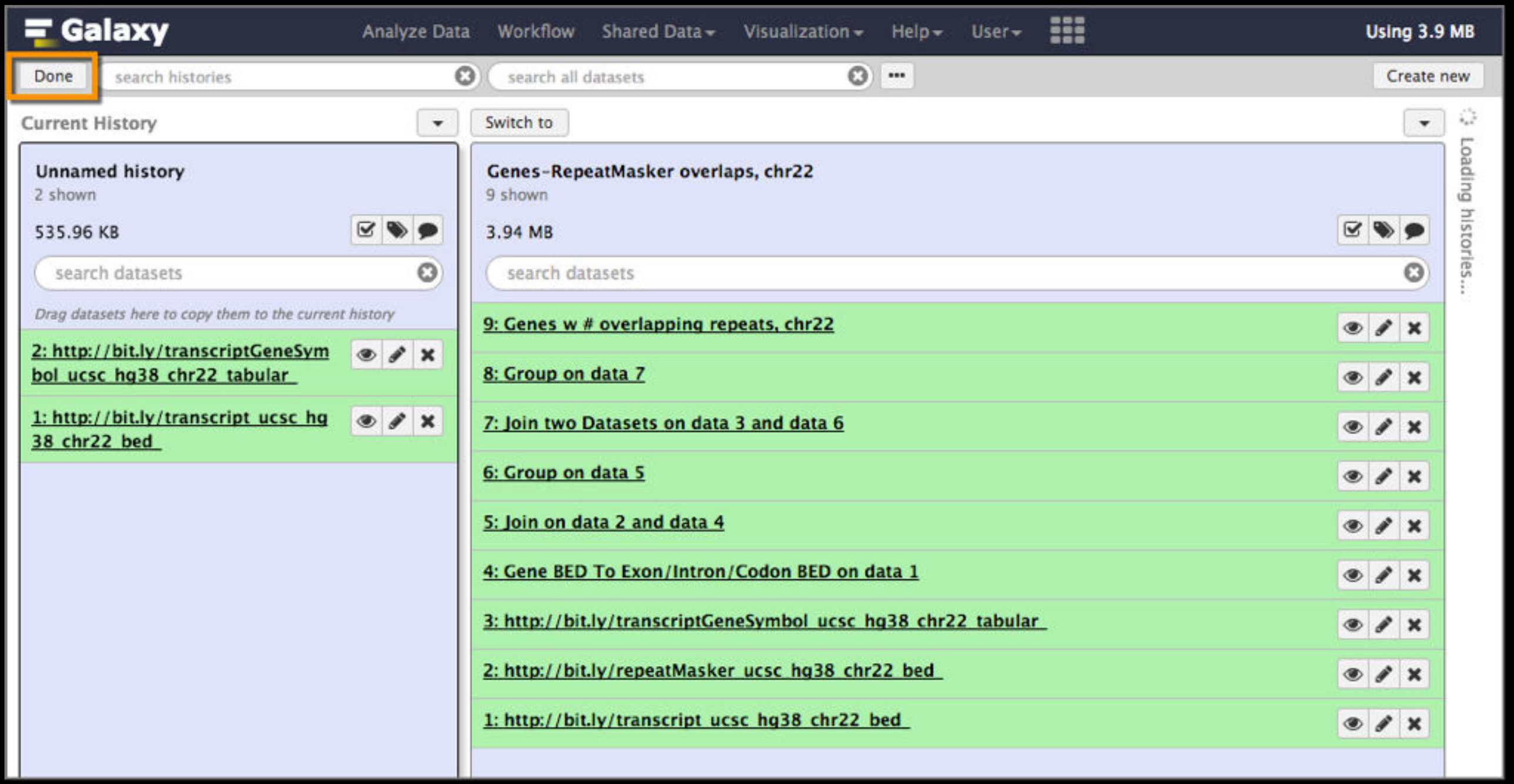


Click the Create New button to create a new history

Drag the transcript and transcript/GeneSymbol datasets from your old history to your new one.



Exit the all histories view

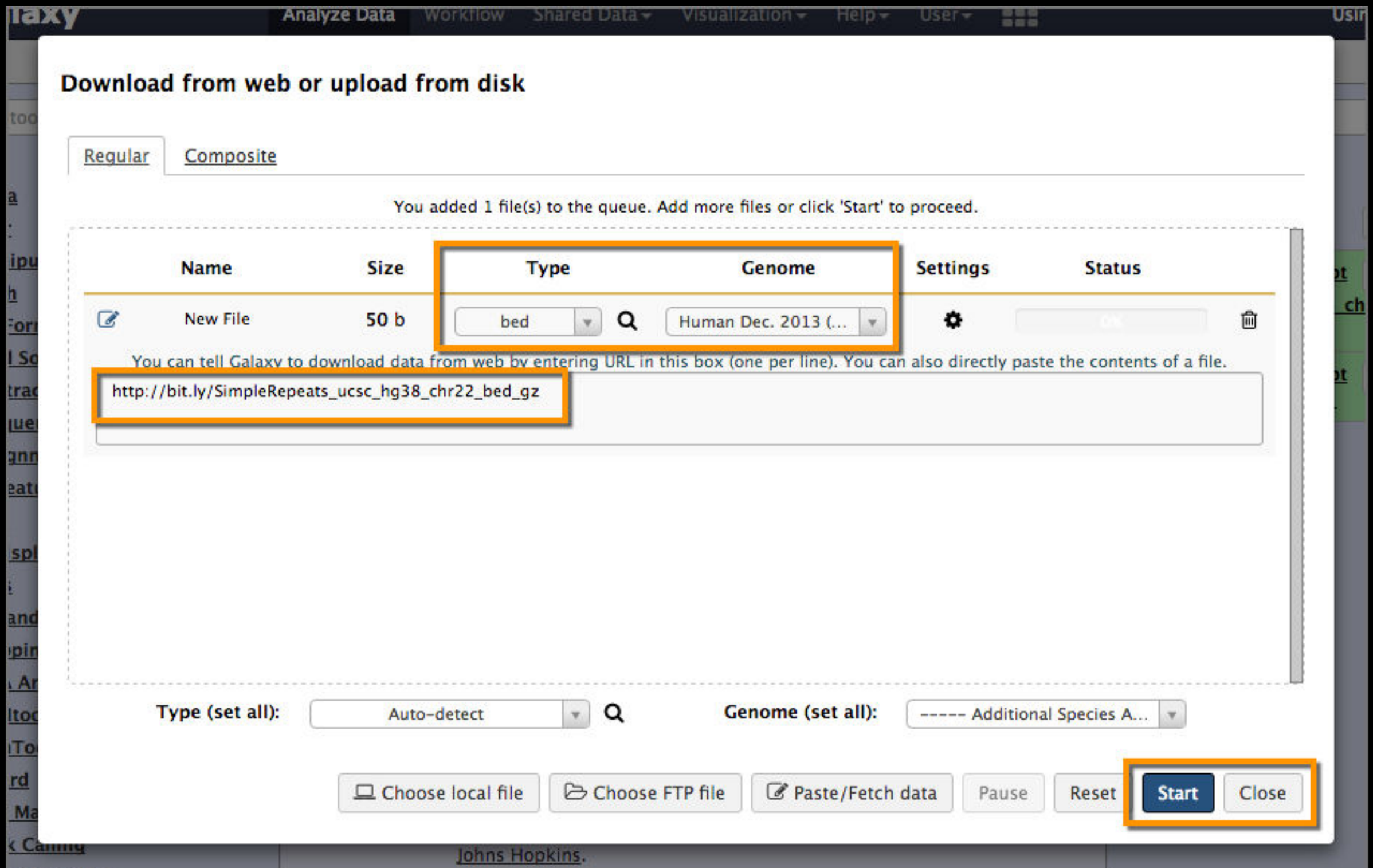


Import TandemRepeatFinder repeats for chr22

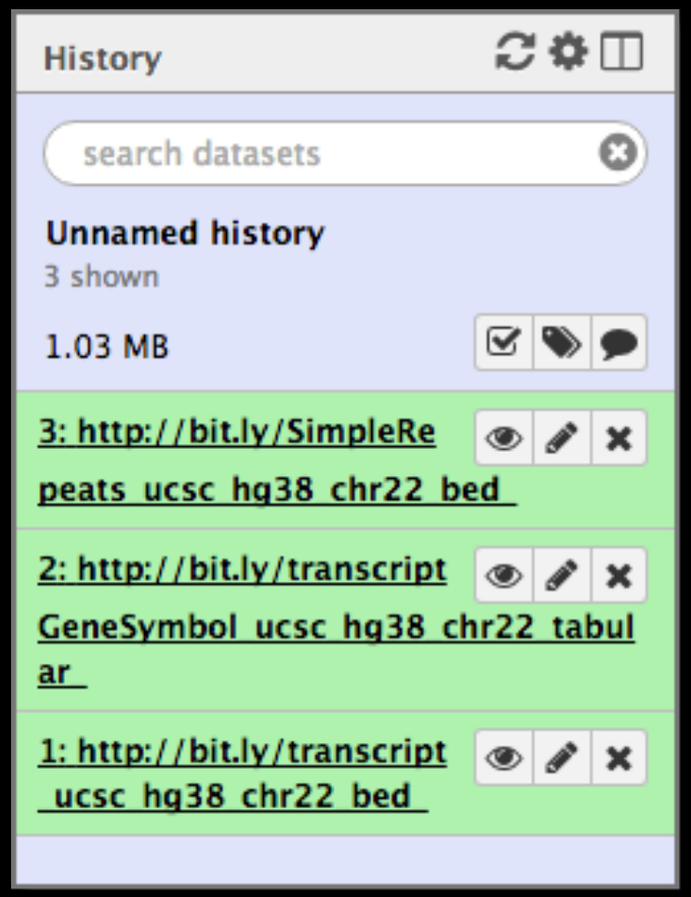
Paste this URL into the paste box:

http://bit.ly/SimpleRepeats_ucsc_hg38_chr22_bed_gz

and set the Type to bed and the Genome to hg38



Our history now has all 3 datasets needed to run the workflow



3. SimpleRepeats for chr22
2. Transcript-Gene Symbol mapping for chr22
1. Transcripts for chr22

Run our workflow using SimpleRepeats

Running workflow "Genes-Repeats overlaps"

Expand All | Collapse

Step 1: Input dataset

12 column BED file defining transcripts. Defines exon structure in columns 10-12.

Transcripts

1: http://bit.ly/transcript_ucsc_hg38_chr22_bed_

type to filter

Step 2: Input dataset

BED file defining repetitive regions.

Repeats

3: http://bit.ly/SimpleRepeats_ucsc_hg38_chr22_bed_

type to filter

Step 3: Input dataset

Two column dataset. First column is the transcript name, 2nd column is the gene the transcript is for.

Transcript - Gene mappings

2: http://bit.ly/transcriptGeneSymbol_ucsc_hg38_chr22_tabular_

type to filter

Step 4: Gene BED To Exon/Intron/Codon BED (version 1.0.0)

Step 5: Join (version 1.0.0)

Step 6: Group (version 2.1.0)

Step 7: Join two Datasets (version 2.0.2)

Step 8: Group (version 2.1.0)

Step 9: Sort (version 1.0.3)

Genes with maximum number of overlapping repeats in any of the gene's transcripts.

Send results to a new history

Run workflow

History

search datasets

Unnamed history

3 shown

1.03 MB

3: http://bit.ly/SimpleRepeats_ucsc_hg38_chr22_bed_

2: http://bit.ly/transcriptGeneSymbol_ucsc_hg38_chr22_tabular_

1: http://bit.ly/transcript_ucsc_hg38_chr22_bed_

Run our workflow using SimpleRepeats

All tasks are queued

Successfully ran workflow "Genes-Repeats overlaps". The following datasets have been added to the queue:

1: http://bit.ly/transcript_ucsc_hg38_chr22_bed_

3: http://bit.ly/SimpleRepeats_ucsc_hg38_chr22_bed_

2: http://bit.ly/transcriptGeneSymbol_ucsc_hg38_chr22_tabular_

4: Gene BED To Exon/Intron/Codon BED on data 1

5: Join on data 3 and data 4

6: Group on data 5

7: Join two Datasets on data 2 and data 6

8: Group on data 7

9: Genes w max # of overlapping repeats

History

search datasets

Unnamed history

9 shown

1.03 MB

9: Genes w max # of overlapping repeats

8: Group on data 7

7: Join two Datasets on data 2 and data 6

6: Group on data 5

5: Join on data 3 and data 4

4: Gene BED To Exon/Intron/Codon BED on data 1

3: http://bit.ly/SimpleRepeats_ucsc_hg38_chr22_bed_

2: http://bit.ly/transcriptGeneSymbol_ucsc_hg38_chr22_tabular_

1: http://bit.ly/transcript_ucsc_hg38_chr22_bed_

All tasks finish. Take a look at the results.

1	2
NEFH	14
MN1	6
C22orf42	5
CARD10	4
MED15	4
TRIOBP	4
ZNF70	4
LMF2	3
SH3BP1	3
SRRD	3
ACR	2
BAIAP2L2	2
BIK	2
CCDC116	2
CCDC157	2
CDC42EP1	2
EWSR1	2
FAM118A	2
MLC1	2
PANX2	2
PHF21B	2
PLXNB2	2
POTEH	2
RANGAP1	2
SREBF2	2
TBC1D10A	2
TUBGCP6	2
ADORA2A	1
APOL3	1
ARVCF	1

History

search datasets

Unnamed history

9 shown

1.61 MB

9: Genes w max # of overlapping repeats

76 lines

format: **tabular**, database: **hg38**

1	2
NEFH	14

8: Group on data 7

7: Join two Datasets on data 2 and data 6

6: Group on data 5

5: Join on data 3 and data 4

4: Gene BED To Exon/Intron/Codon BED on data 1

3: http://bit.ly/SimpleRepeats_ucsc_hg38_chr22_bed

2: http://bit.ly/transcriptGeneSymbol_ucsc_hg38_chr22_tabular

And finally give the output dataset and the history meaningful names

Time allowing

Sharing and publishing
Exporting
Comparing Gene Lists

Agenda

Launching Galaxy in Jetstream

A quick stroll through the Galaxy

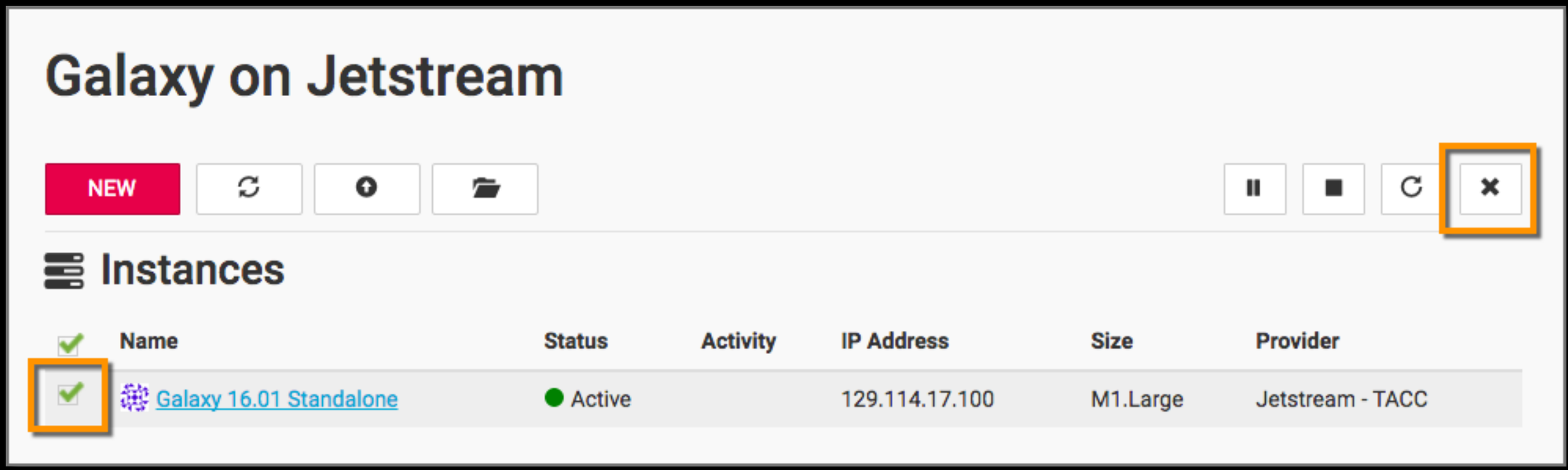
Demonstrate Galaxy by addressing a specific question

Turn that analysis into a reusable workflow

Shutting down Galaxy in Jetstream

Galaxy Ecosystem

bit.ly/ww_gxy_slides



Select the running Galaxy instance and
then click the delete (x) icon

Agenda

Launching Galaxy in Jetstream

A quick stroll through the Galaxy

Demonstrate Galaxy by addressing a specific question

Turn that analysis into a reusable workflow

Shutting down Galaxy in Jetstream

Galaxy Ecosystem

bit.ly/ww_gxy_slides

2016 Galaxy Community Conference (GCC2016)

June 25-29, 2016
Bloomington, Indiana
galaxyproject.org/GCC2016

Slides & posters are now online. Video will be shortly

26 - 30 June France

GCC 2017 Montpellier

Le Corum
Conference centre

gcc2017.sciencesconf.org

November 7-11

2016

GALAXY

ADMIN TRAINING

SLC

Salt Lake City, Utah

#GAMe2017
@EMBL_ABR

GAMe 2017

Galaxy Australasia Meeting

3 - 9 February, Melbourne, Australia

www.embl-abr.org.au/GAMe2017

Galaxy Community Resources: Galaxy Biostar

Tens of thousands of users leads to a lot of questions.

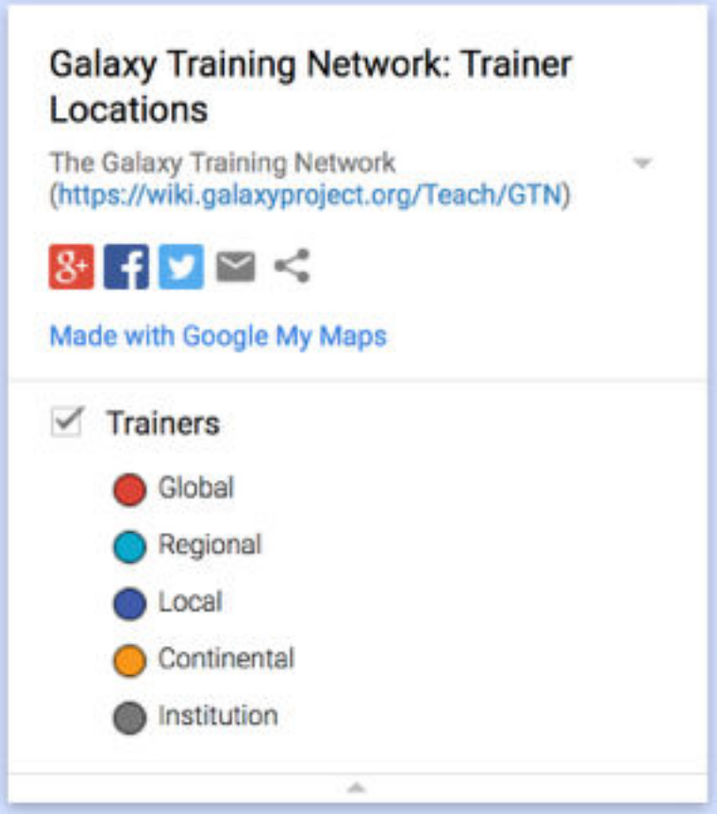
Absolutely have to encourage community support.

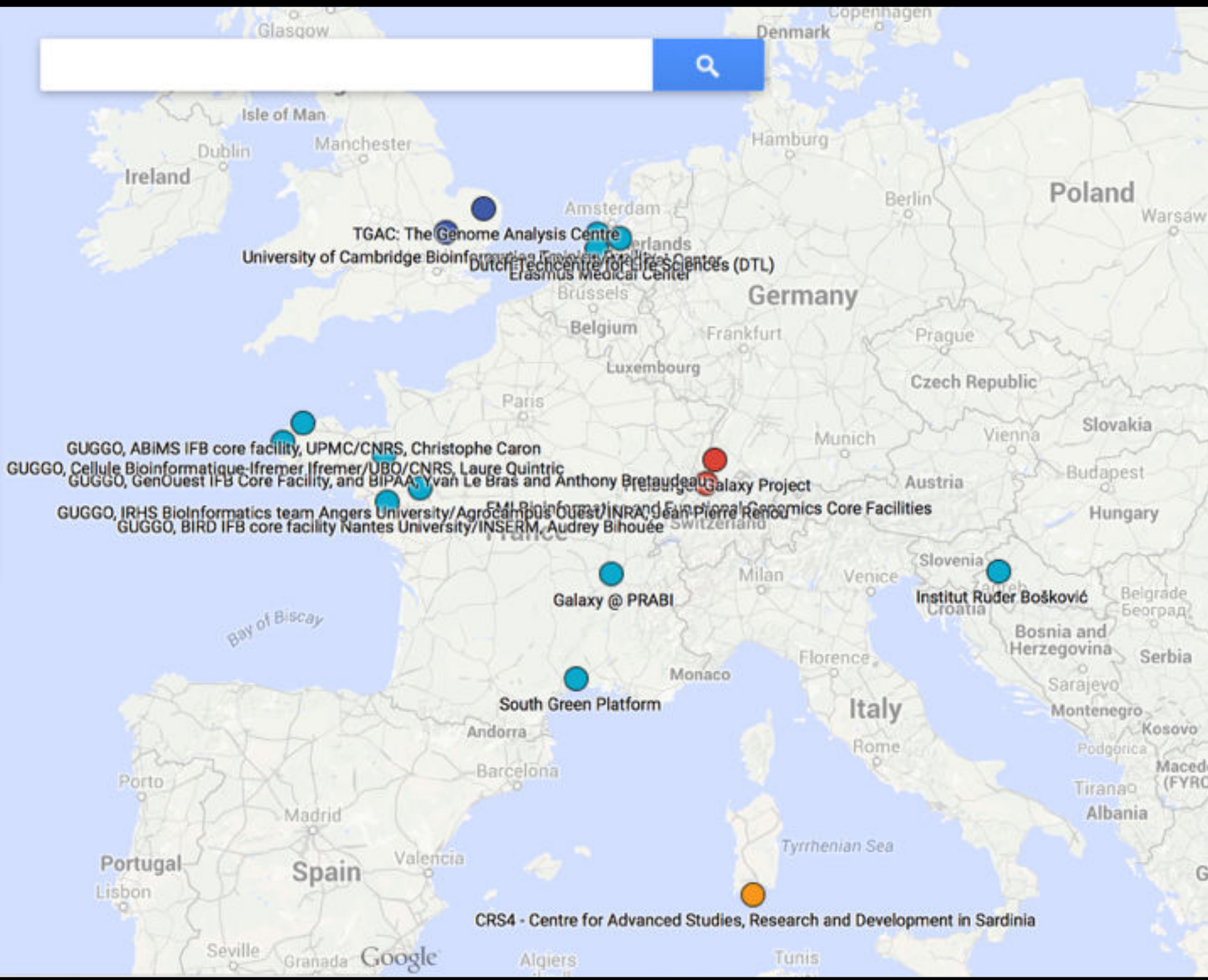
Project traditionally used mailing list

Moved the user support list to Galaxy Biostar, an online forum, that uses the Biostar platform

https://biostar.usegalaxy.org/

Scaling Training





Galaxy Training Network

bit.ly/gxygtn

The Galaxy Team

Enis Afgan

Dannon Baker

Dan Blankenberg

Dave Bouvier

Marten Cech

John Chilton

Dave Clements

Nate Coraor

Carl Eberhard

Jeremy Goecks

Sam Guerler

Jen Jackson

Ross Lazarus

Anton Nekrutenko

Nick Stoler

James Taylor

Nitesh Turaga

http://wiki.galaxyproject.org/GalaxyTeam

Acknowledgements

You

Scott Lathrop
Michael Miller
Mike Pingleton
XSEDE
Jetstream

NIH
Johns Hopkins University
Penn State University

Thanks