

Galaxy Workflows

Workflows Workshop

August 9-10, 2016

Online

Dave Clements & Enis Afgan

Galaxy Team

Johns Hopkins University

<http://galaxyproject.org/>

The logo for XSEDE, consisting of the letters 'XSEDE' in a bold, blue, sans-serif font.

Extreme Science and Engineering
Discovery Environment



#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 is available several ways ...

bit.ly/ww_gxy_slides

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

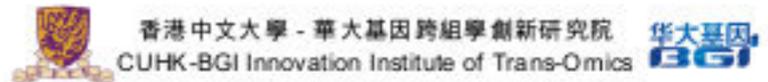
Galaxy / Metabiome Portal



The Microbiome Analysis Center
Life on a Smaller Scale

Welcome to the Metabiome Portal @ GMU

We have developed the MAC Metabiome Portal, a flexible and extensible web browser, with the aim of simplifying access, usage, storage, and analysis of microbiome and metagenomic data. The Portal uses a robust database management system and offers a range of analytical tools and visualization options, such as taxonomic clustering, network analysis, and phylogenetic analysis.



(GIGA)ⁿ Galaxy
by CBIIT

Integrated publishing of workflows from GIGAⁿ SCIENCE

Cistrome



A Galaxy Server dedicated to ChIP-* analysis



Public Galaxy Servers
and *still* counting



The Genomic HyperBrowser

Powered by Galaxy

SCDE STEM CELL DISCOVERY ENGINE



Experiments Connected



Whale Shark Galaxy! 

South Green
bioinformatics platform

Genomic analysis tools for southern and Mediterranean plants

bit.ly/gxyServers

Galaxy is available on the Cloud

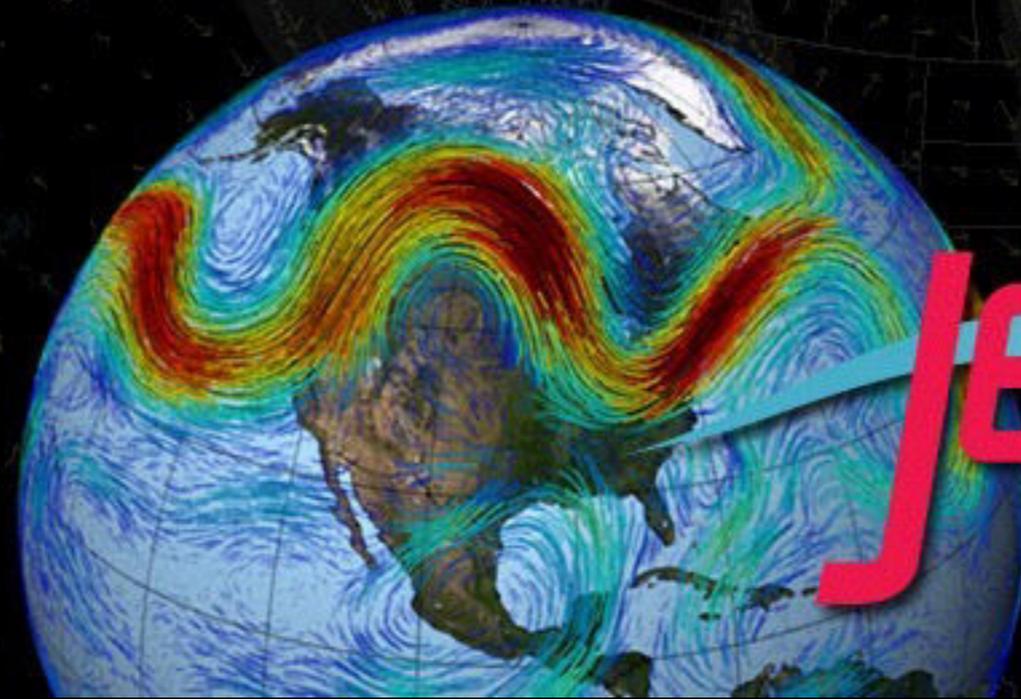


<http://aws.amazon.com/education>

<http://globus.org/>

<http://wiki.galaxyproject.org/Cloud>

Via



Jetstream

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?

Genes: Exons, Introns, and alternative versions

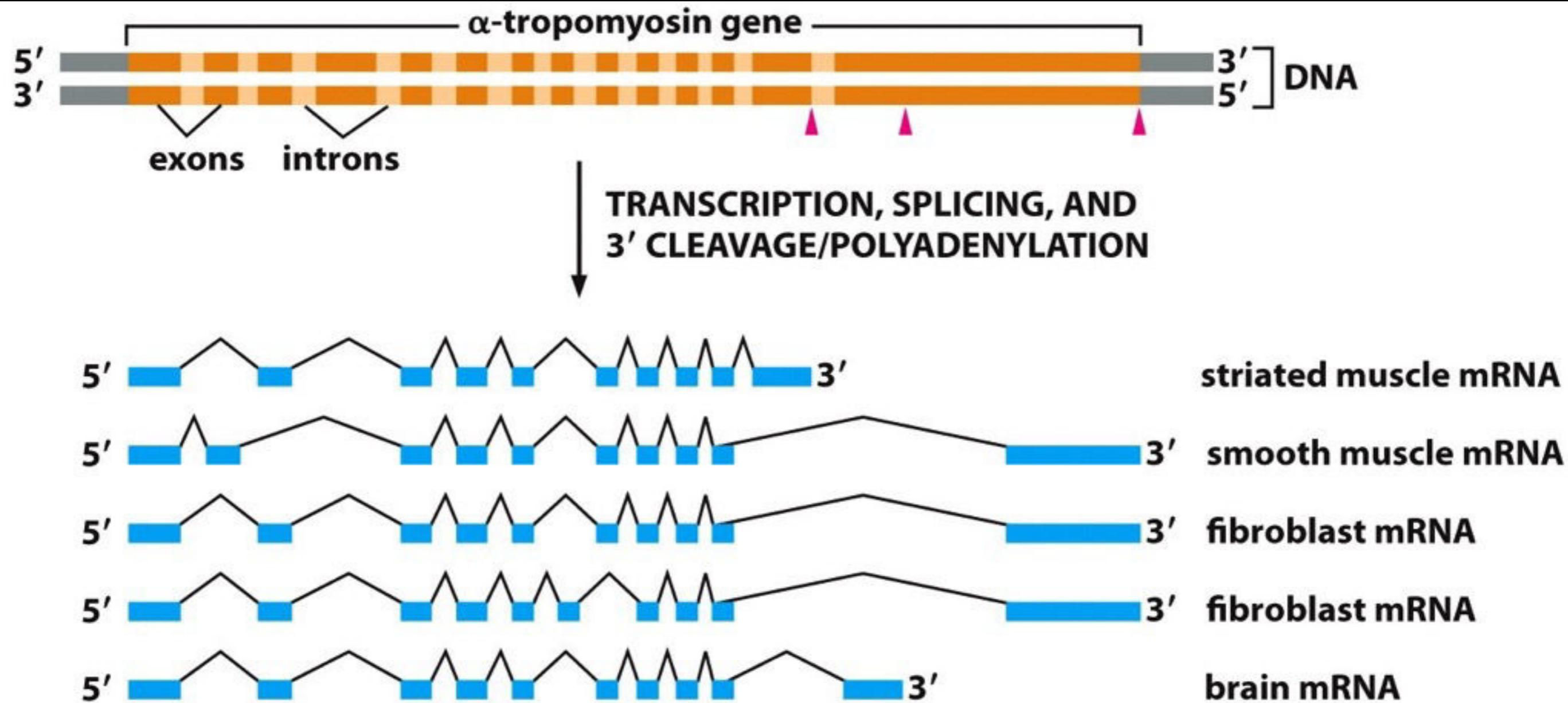


Figure 7-21 Essential Cell Biology 3/e (© Garland Science 2010)

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:

Galaxy on Jetstream

NEW  

Instances

<input type="checkbox"/>	Name	Status	Activity	IP Address	Size	Provider
<input type="checkbox"/>	 Galaxy 16.01 Standalone	 Build	Spawning	N/A	M1.Large	Jetstream - TACC

<input type="checkbox"/>	Name	Status	Activity	IP Address	Size	Provider
<input type="checkbox"/>	 Galaxy 16.01 Standalone	 Active	Deploying	129.114.17.71	M1.Large	Jetstream - TACC

<input type="checkbox"/>	Name	Status	Activity	IP Address	Size	Provider
<input type="checkbox"/>	 Galaxy 16.01 Standalone	 Active		129.114.17.71	M1.Large	Jetstream - TACC

Copy & paste IP address into a new browser tab.



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: Comini](#)

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](#).



search datasets



Unnamed history

0 b

i This history is empty. You can [load your own data](#) or [get data from an external source](#)

Create a login on your server

The image shows a web application interface. At the top left, there is a 'User' dropdown menu with options for 'Login' and 'Register'. Below this is a 'Create account' form with the following fields: 'Email address:', 'Password:', 'Confirm password:', and 'Public name:'. Below the 'Public name' field, there is a note: 'Your public name is an identifier that will be used to generate addresses for information you share publicly. Public names must be at least three characters in length and contain only lower-case letters, numbers, and the '-' character.' A 'Submit' button is located at the bottom of the form. Below the form, there is a navigation bar with links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. At the bottom, a green notification box displays a checkmark and the text: 'Now logged in as clements@galaxyproject.org. [Return to the home page.](#)'

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

Get Data: Genes and Repeats

Import 3 data files into your Galaxy instance

The image shows a screenshot of the Galaxy web interface. On the left, the 'Tools' menu is open, with the 'Get Data' option highlighted. The main area of the interface is a large white box with a dashed border, containing the text 'Drop files here' and a folder icon. Below this box, there are several controls: 'Type (set all):' with a dropdown menu set to 'Auto-detect', a search icon, 'Genome (set all):' with a dropdown menu set to '----- Additional Species A...', and a row of buttons: 'Choose local file', 'Choose FTP file', 'Paste/Fetch data' (highlighted with an orange box), 'Pause', 'Reset', 'Start', and 'Close'.

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

You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
 New File	97 b	Auto-dete... 	 ----- Additional Sp.. 		<input type="text" value="0%"/> 

You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.

```
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)

Name	Size	Type	Genome	Settings
New File	159 b	Auto-dete... <input type="text"/>	----- Additional Sp.. <input type="text"/>	<input type="checkbox"/>
You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the content of a file into this box.			hg38	<input type="checkbox"/>
http://bit.ly/transcript_ucsc_hg38_chr22_bed_gz			Human Dec. 2013 (GRCh38/hg38) (hg38)	
http://bit.ly/repeatMasker_ucsc_hg38_chr22_bed_gz				

Set the data **type** to **bed**

Name	Size	Type	Genome	Settings
New File	97 b	Auto-dete... <input type="text"/>	----- Additional Sp.. <input type="text"/>	<input type="checkbox"/>
You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the content of a file into this box.				
http://bit.ly/transcript_ucsc_hg38_chr22_bed_gz				
http://bit.ly/repeatMasker_ucsc_hg38_chr22_bed_gz				

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

You added 2 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
 New File	97 b	bed 	 ----- Additional Sp.. 		
You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.					
<pre>http://bit.ly/transcript_ucsc_hg38_chr22_bed_gz http://bit.ly/repeatMasker_ucsc_hg38_chr22_bed_gz</pre>					
 New File	61 b	Auto-dete... 	 ----- Additional Sp.. 		
You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.					
<pre>http://bit.ly/transcriptGeneSymbol_ucsc_hg38_chr22_tabular_gz</pre>					

Type (set all):  Genome (set all): 

Get Data: Genes and Repeats

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

New File 61 b tabular Human Dec. 2013 (...) 0%

You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.

http://bit.ly/transcriptGeneSymbol_ucsc_hg38_chr22_tabular_gz

Then click **Start**, and then **Close**.

Type (set all): Auto-detect Genome (set all): ----- Additional Species A...

Choose local file Choose FTP file Paste/Fetch data Pause Reset Start Close

Sciences, The Institute for CyberScience at Penn

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.

Using 0 bytes

History  

search datasets 

Unnamed history
3 shown

0 b 

3: http://bit.ly/transcriptGeneSymbol_ucsc_hg38_chr22_tabular_gz   

2: http://bit.ly/repeatMasker_ucsc_hg38_chr22_bed_gz   

1: http://bit.ly/transcript_ucsc_hg38_chr22_bed_gz   

Using 3.3 MB

History  

search datasets 

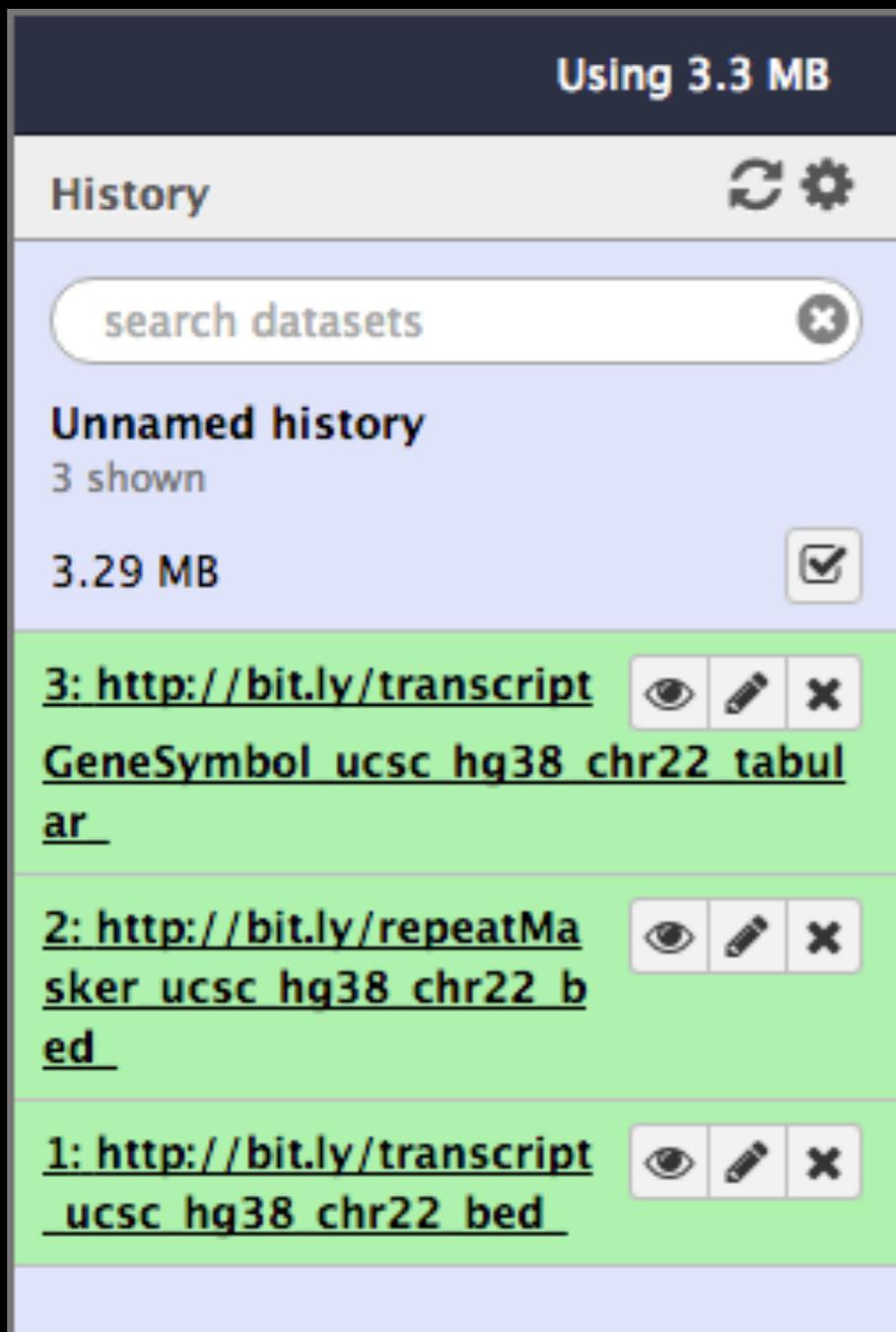
Unnamed history
3 shown

3.29 MB 

3: http://bit.ly/transcriptGeneSymbol_ucsc_hg38_chr22_tabular_gz   

2: http://bit.ly/repeatMasker_ucsc_hg38_chr22_bed_gz   

1: http://bit.ly/transcript_ucsc_hg38_chr22_bed_gz   



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

Using 3.3 MB

History  

search datasets 

Unnamed history
3 shown

3.29 MB 

3: <http://bit.ly/transcript>   
[GeneSymbol ucsc hg38 chr22 tabular](#)

2: http://bit.ly/repeatMasker_ucsc_hg38_chr22_bed   

1: http://bit.ly/transcript_ucsc_hg38_chr22_bed   

4,093 regions
format: **bed**, database: **hg38**

uploaded bed file

display in IGB [View](#)
display with IGV [local](#) [Human hg38](#)

1. Chrom	2. Start	3. End	4. Name
chr22	10736170	10736283	uc062bdo.1

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	+

History   

search datasets 

Unnamed history
3 shown

3.29 MB   

3: <http://bit.ly/transcript>   
[GeneSymbol ucsc hg38 chr22 tabular](#)

2: http://bit.ly/repeatMasker_ucsc_hg38_chr22_bed   

1: http://bit.ly/transcript_ucsc_hg38_chr22_bed   
4,093 regions
format: **bed**, database: **hg38**

uploaded bed file     

display in IGB [View](#)
display with IGV [local](#) [Human](#) [hg38](#)

1.Chrom	2.Start	3.End	4.Name
chr22	10736170	10736283	uc062bdo.1

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

Our first tool: Extract exons from genes

The screenshot shows the Galaxy web interface. On the left is a 'Tools' panel with a search bar and a list of tool categories. The 'Extract Features' category is highlighted with an orange box, and the tool 'Gene BED To Exon/Intron/Codon BED expander' is selected. The main panel shows the tool's configuration page. The 'Extract' dropdown is set to 'Coding Exons only'. The 'from' dropdown is set to '1: http://bit.ly/transcript_ucsc_hg38_chr22_bed_'. The 'Execute' button is highlighted with an orange box. A warning message states: 'This tool works only on a BED file that contains at least 12 fields (see Example and About formats below). The output will be empty if applied to a BED file with 3 or 6 fields.' Below this is a 'What it does' section explaining that the tool unpacks gene information into individual exons, introns, and UTRs.

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	+

History   

search datasets 

Unnamed history
4 shown

3.84 MB   

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

14,875 regions
format: **bed**, database: **hg38**

Region: coding;

display in IGB [View](#)
display with IGV [local](#) [Human hg38](#)

1.Chrom	2.Start	3.End	4.Name
chr22	11065973	11066015	uc062bdq.1

3: <http://bit.ly/transcrip>   

6 column dataset, 1 record per exon

Transcript name has become the **exon name**

Identify which **exons** have overlapping repeats

The screenshot shows the Galaxy web interface. On the left, the 'Tools' panel is open, and the 'Operate on Genomic Intervals' toolbox is selected. The 'Join the intervals of two datasets side-by-side' tool is highlighted. The main panel shows the tool's configuration page, which includes the following fields:

- Join:** A dropdown menu with the selected option '4: Gene BED To Exon/Intron/Codon BED on data 1'.
- First dataset:** A text input field containing the selected dataset ID.
- with:** A dropdown menu with the selected option '2: http://bit.ly/repeatMasker_ucsc_hg38_chr22_bed_'.
- Second dataset:** A text input field containing the selected dataset ID.
- with min overlap:** A text input field containing the value '1'.
- (bp):** A label indicating the unit of the overlap value.
- Return:** A dropdown menu with the selected option 'Only records that are joined (INNER JOIN)'.
- Execute:** A blue button with a checkmark icon and the text 'Execute'.

Below the configuration fields, there is a tip: **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.

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	-

History   

search datasets 

Unnamed history
5 shown

3.91 MB   

5: Join on data 2 and data 4   

911 regions
format: interval, database: hg38

display with IGV [local](#) [Human](#)
[hg38](#)

1. Chrom	2. Start	3. End	4. Name
chr22	11065973	11066015	uc062bdq.1

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

3: http://bit.ly/transcriptGeneSymbol_ucsc_hg38_chr   

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

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

The screenshot shows the Galaxy web interface. On the left is a 'Tools' sidebar with a search bar and a list of tool categories. The 'Join, Subtract and Group' category is highlighted with an orange box. Below it, the 'Group data by a column and perform aggregate operation on other columns.' tool is also highlighted with an orange box. The main panel on the right shows the tool's configuration options. The 'Select data' dropdown is set to '5: Join on data 2 and data 4'. The 'Group by column' dropdown is set to 'Column: 4' and is highlighted with an orange box. Below this, there are options for 'Ignore case while grouping?' (Yes/No) and 'Ignore lines beginning with these characters' (checkboxes for >, @, +, <, *, -).

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

5: Join on data 2 and data 4

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

>

@

+

<

*

-

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

^
 #

lines beginning with these are not grouped

Operation

+ Insert Operation

✓ Execute

^
 #

lines beginning with these are not grouped

Operation

1: Operation 

Type

Count

On column

Column: 1

Round result to nearest integer?

NO

+ Insert Operation

✓ Execute

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
uc062ben.1	LINC01297

History   

search datasets 

Unnamed history
6 shown
3.92 MB   

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

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

2: http://bit.ly/repeatMasker_ucsc_hg38_chr22_bed   

1: http://bit.ly/transcript_ucsc_hg38_chr22_bed   

Associate transcript counts with genes

The screenshot displays the Galaxy web interface. On the left is a 'Tools' sidebar with a search bar and a list of tool categories. The 'Join, Subtract and Group' category is highlighted with an orange box. Within this category, the tool 'Join two Datasets side by side on a specified field' is also highlighted with an orange box. The main panel shows the configuration for this tool. The title is 'Join two Datasets side by side on a specified field (Galaxy Version 2.0.2)'. The 'Join' section has a dropdown menu set to '6: Group on data 5'. The 'using column' dropdown is set to 'Column: 1'. The 'with' section has a dropdown menu set to '3: http://bit.ly/transcriptGeneSymbol_ucsc_hg38_chr22_t...'. The 'and column' dropdown is set to 'Column: 1'. There are three checkboxes, all set to 'No': 'Keep lines of first input that do not join with second input', 'Keep lines of first input that are incomplete', and 'Fill empty columns'. At the bottom, the 'Execute' button is highlighted with an orange box.

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

History   

search datasets 

Unnamed history
7 shown
3.93 MB   

7: Join two Datasets on data 3 and data 6   
628 lines
format: tabular, database: hg38
     
1 2 3 4
uc002zly.5 2 uc002zly.5 IL17RA

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/transcript>   

For some genes, we have different overlap counts.

For each gene, get the max # of overlaps

The screenshot shows the Galaxy web interface with the 'Join, Subtract and Group' tool selected. The tool's configuration is as follows:

- Group data by a column and perform aggregate operation on other columns.** (Galaxy Version 2.1.0)
- Select data:** 7: Join two Datasets on data 3 and data 6
- Group by column:** Column: 4
- Ignore case while grouping?** Yes
- Ignore lines beginning with these characters:** Select/Unselect all
- Operation:** 1: Operation
 - Type:** Maximum
 - On column:** Column: 2
 - Round result to nearest integer?** NO
- Execute** button

The tool description in the left sidebar is: "Group data by a column and perform aggregate operation on other columns."

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

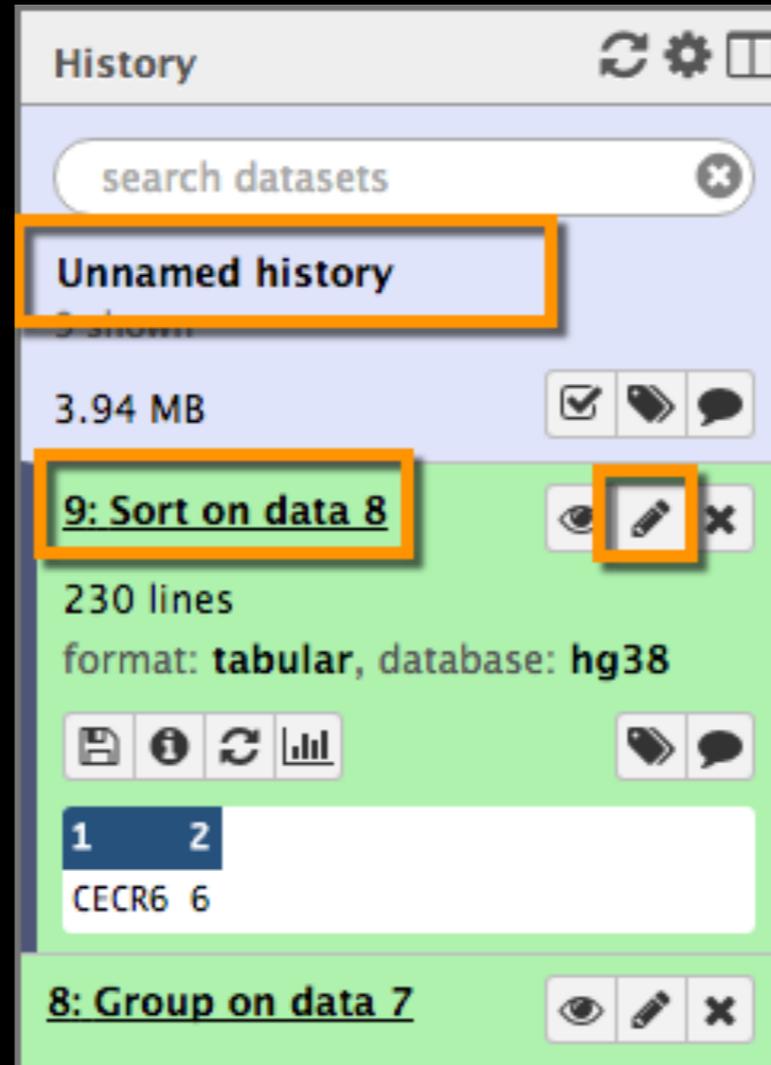
Execute

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
CFLSR1	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, chr22

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

The screenshot shows a 'History' panel on the right side of a web interface. The panel contains a list of history items, each with a number and a description. A dropdown menu is open over the panel, listing various actions. The 'Extract Workflow' option is highlighted with an orange box. The menu is organized into sections: HISTORY LISTS, HISTORY ACTIONS, DATASET ACTIONS, DOWNLOADS, and OTHER ACTIONS. The 'Extract Workflow' option is located under the HISTORY ACTIONS section.

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

At the bottom of the history panel, there is a list of items with a URL: [1: http://bit.ly/transcript_ucsc_hg38_chr22_bed](http://bit.ly/transcript_ucsc_hg38_chr22_bed). To the right of the URL are three icons: an eye, a pencil, and an 'X'.

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:

Tool	History items created
<input type="button" value="Upload File"/> <i>This tool cannot be used in workflows</i>	1: http://bit.ly/transcript_ucsc_hq38_chr22_bed <input checked="" type="checkbox"/> Treat as input dataset
<input type="button" value="Upload File"/> <i>This tool cannot be used in workflows</i>	2: http://bit.ly/repeatMasker_ucsc_hq38_chr22_bed <input checked="" type="checkbox"/> Treat as input dataset
<input type="button" value="Upload File"/> <i>This tool cannot be used in workflows</i>	3: http://bit.ly/transcriptGeneSymbol_ucsc_hq38_chr22_tabular <input checked="" type="checkbox"/> Treat as input dataset
<input type="button" value="Gene BED To Exon/Intron/Codon BED"/> <input checked="" type="checkbox"/> Include "Gene BED To Exon/Intron/Codon BED" in workflow	4: Gene BED To Exon/Intron/Codon BED on data 1
<input type="button" value="Join"/> <input checked="" type="checkbox"/> Include "Join" in workflow	5: Join on data 2 and data 4
<input type="button" value="Group"/> <input checked="" type="checkbox"/> Include "Group" in workflow	6: Group on data 5
<input type="button" value="Join two Datasets"/> <input checked="" type="checkbox"/> Include "Join two Datasets" in workflow	7: Join two Datasets on data 3 and data 6
<input type="button" value="Group"/> <input checked="" type="checkbox"/> Include "Group" in workflow	8: Group on data 7
<input type="button" value="Sort"/> <input checked="" type="checkbox"/> 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_hq38_chr22_tabular
- 2:** http://bit.ly/repeatMasker_ucsc_hq38_chr22_bed
- 1:** http://bit.ly/transcript_ucsc_hq38_chr22_bed

Edit the workflow

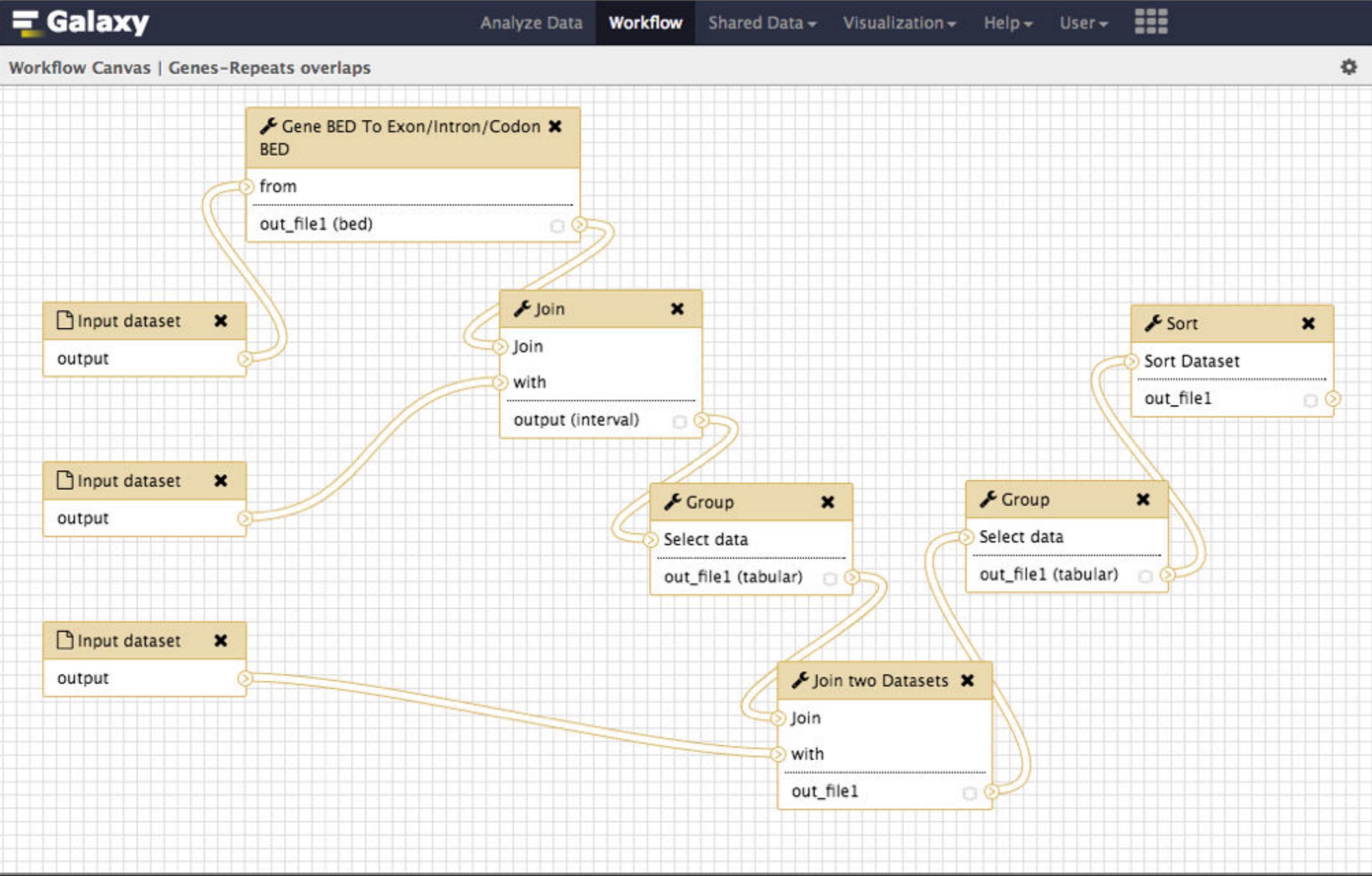


Workflow "Genes-Repeats overlaps" created from current history. You can [edit](#) or [run](#) the workflow.

The screenshot displays the Galaxy workflow editor interface. The top navigation bar includes "Galaxy", "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", and "User". The main area is the "Workflow Canvas" for a workflow named "Genes-Repeats overlaps". The canvas shows a sequence of tools: three "Input dataset" tools, a "Gene BED To Exon/Intron/Codon" tool, a "Join" tool, a "Group" tool, a "Join two Datasets" tool, another "Group" tool, and a "Sort" tool. The workflow is connected by yellow lines. On the right, the "Details" panel is visible, containing "Edit Workflow Attributes" with fields for "Name" (Genes-Repeats overlaps), "Tags", and "Annotation / Notes".

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

Workflow Canvas | Genes-Repeats overlaps

Gene BED To Exon/Intron/Codon BED
BED
from
out_file1 (bed)

Join
Join
with
output (interval)

Group
Select data
out_file1 (tabular)

Group
Select data
out_file1 (tab

Join two Datasets
Join

Input dataset
output

Input dataset
output

Input dataset
output

Details

Input dataset

Name:
Transcripts

Edit Step Attributes

Annotation / Notes:
12 column BED file defining transcripts. Defines exon structure in columns 10-12.

Add an annotation or notes to this step; annotations are available when a workflow is viewed.

Name the transcripts input dataset.

Give the input and output datasets meaningful names

Workflow Canvas | Genes-Repeats overlaps

Details

Input dataset

Name:
Repeats

Edit Step Attributes

Annotation / Notes:
BED file defining repetitive regions.
Add an annotation or notes to this step; annotations are available when a workflow is viewed.

Name the repeats input dataset.

Give the input and output datasets meaningful names

Workflow Canvas | Genes-Repeats overlaps

Gene BED To Exon/Intron/Codon BED
BED
from
out_file1 (bed)

Input dataset
output

Join
Join
with
output (interval)

Group
Select data
out_file1 (tabular)

Group
Select data
out_file1 (tab

Join two Datasets
Join
with

Input dataset
output

Input dataset
output

Details

Input dataset

Name:
Transcript - Gene mappings

Edit Step Attributes

Annotation / Notes:
Two column dataset. First column is the transcript name, 2nd column is the gene the transcript is for.
Add an annotation or notes to this step; annotations are available when a workflow is viewed.

Name the Transcript Gene mapping input dataset.

Name the output dataset.

Workflow Canvas | Genes-Repeats overlaps

The workflow consists of the following steps:

- BED To Exon/Intron/Codon**: Input (bed)
- Join**: Join with output (interval)
- Group**: Select data out_file1 (tabular)
- Sort Dataset**: Sort Dataset out_file1 (highlighted with an orange box)
- Group**: Select data out_file1 (tabular)
- Join two Datasets**: Join with 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.

Add an annotation or note for this 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

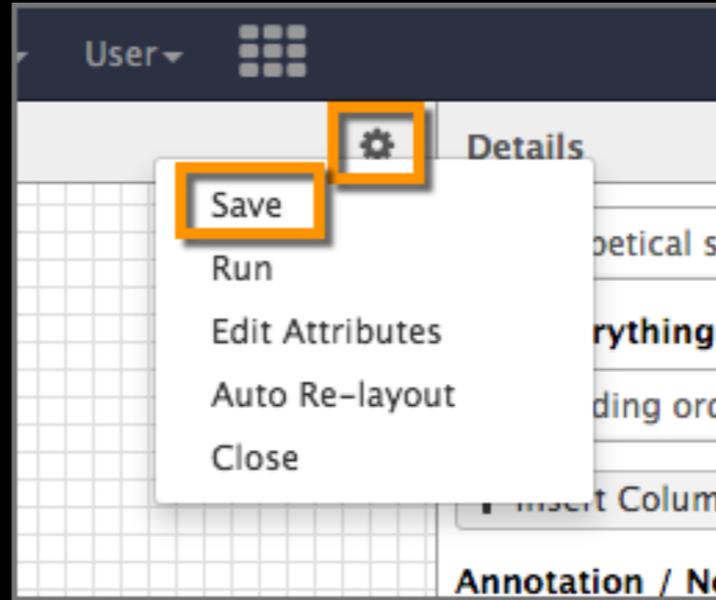
This will provide a short name to describe the output - this must be unique across workflows.

Rename dataset

Genes w max # of overlapping repe

This action will rename the output dataset. Click here for more

Save your workflow edits!



Review your workflows

Galaxy Analyze Data **Workflow** Shared Data Visualization Help User Using 3.9 MB

Your workflows

[+ Create new workflow](#) [↑ Upload or import workflow](#)

Name	# of Steps
Genes-Repeats overlaps ▾	9

Workflows shared with you by others

No workflows have been shared with you.

Other options

[Configure your workflow menu](#)

Start a new history

The screenshot shows the Galaxy web interface. The top navigation bar includes the Galaxy logo, the 'Analyze Data' tab (highlighted with an orange box), and other tabs like 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The main content area displays a welcome message: 'Welcome to Galaxy on the Jetstream Cloud' and 'Galaxy on the Jetstream Cloud is ready for use!'. On the right side, the 'History' panel is visible, showing a list of datasets. The 'View all histories' icon (a grid of three squares) in the top right of the history panel is highlighted with an orange box.

Galaxy

Analyze Data Workflow Shared Data Visualization Help User Using 3.9 MB

Tools

search tools

Get Data
Send Data
Lift-Over
Text Manipulation
Datamash
Convert Formats
Filter and Sort
Join, Subtract and Group
Fetch Sequences
Fetch Alignments

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

Galaxy on the Jetstream Cloud is ready for use!

To learn how to use Galaxy please see the wiki

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

Start a new history

The screenshot shows the Galaxy web interface. At the top, there is a navigation bar with the Galaxy logo and several menu items: Analyze Data, Workflow, Shared Data, Visualization, Help, and User. In the top right corner, it says 'Using 3.9 MB'. Below the navigation bar, there are search bars for 'search histories' and 'search all datasets'. A 'Create new' button is highlighted with an orange box in the top right corner. The main content area is divided into two panels. The left panel, titled 'Current History', shows an 'Unnamed history' with '0 b' of data and a message that says 'This history is empty'. The right panel, titled 'Switch to', shows a list of 9 datasets under the heading 'Genes-RepeatMasker overlaps, chr22'. The datasets are numbered 1 through 9, with 1 being the base data and 9 being the final result. Each dataset entry has a search bar and a set of icons (eye, pencil, and X) for viewing, editing, and deleting the dataset.

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.

Galaxy Analyze Data Workflow Shared Data Visualization Help User Using 3.9 MB

Done search histories search all datasets Create new

Current History Unnamed history 1 shown 462.45 KB search datasets Drag datasets here to copy them to the current history

Switch to Genes-RepeatMasker overlaps, chr22 9 shown 3.94 MB search datasets Loading histories...

Current History Dataset:
1: <http://bit.ly/transcript ucsc hg38 chr22 bed>

Switch to Workflow Steps:
9: [Genes w # overlapping repeats, chr22](#)
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>

Exit the all histories view

Galaxy Analyze Data Workflow Shared Data Visualization Help User Using 3.9 MB

Done search histories search all datasets Create new

Current History Switch to

Unnamed history
2 shown
535.96 KB

search datasets

Drag datasets here to copy them to the current history

2: http://bit.ly/transcriptGeneSymbol_ucsc_hg38_chr22_tabular

1: http://bit.ly/transcript_ucsc_hg38_chr22_bed

Genes-RepeatMasker overlaps, chr22
9 shown
3.94 MB

search datasets

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

Loading histories...

Import TandemRepeatFinder repeats for chr22

The screenshot displays the Galaxy web interface. The top navigation bar includes the Galaxy logo, menu items for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User', and a status indicator 'Using 3.9 MB'. The 'Tools' panel on the left contains a search bar and a list of tool categories: 'Get Data', 'Send Data', 'Lift-Over', 'Text Manipulation', 'Datamash', 'Convert Formats', 'Filter and Sort', 'Join, Subtract and Group', and 'Fetch Sequences'. An orange box highlights the upload icon in the 'Tools' panel header. The main workspace shows a 'Welcome to Galaxy on the Jetstream Cloud' message. The 'History' panel on the right shows two entries: '2: http://bit.ly/transcript ucsc hg38 chr22 tabular' and '1: http://bit.ly/transcript ucsc hg38 chr22 bed', both highlighted in green.

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**

Download from web or upload from disk

Regular Composite

You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

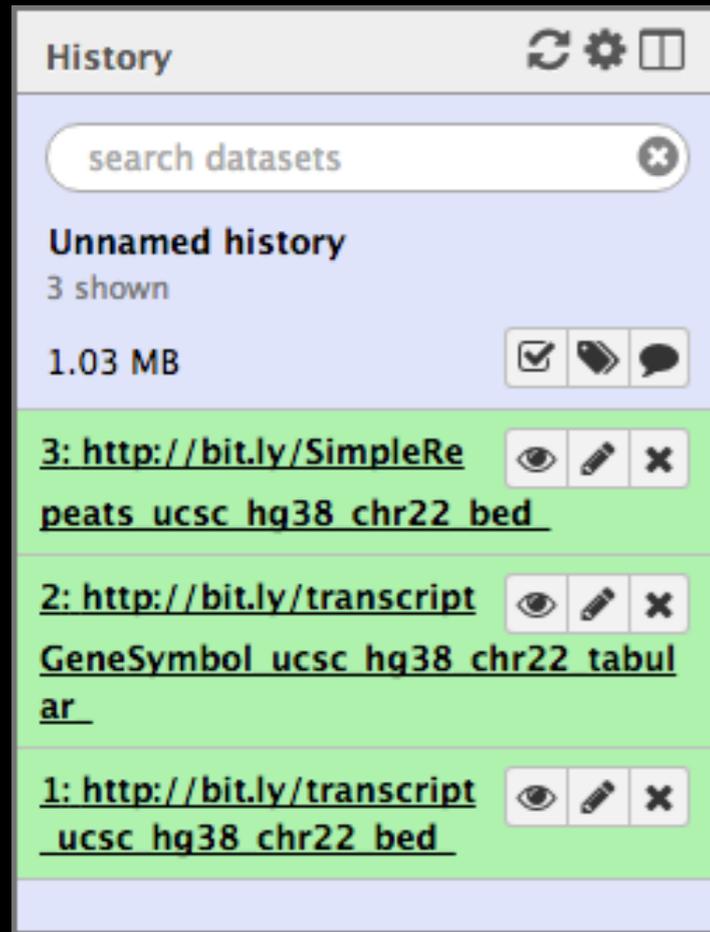
Name	Size	Type	Genome	Settings	Status
New File	50 b	bed	Human Dec. 2013 (...)		

You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.

http://bit.ly/SimpleRepeats_ucsc_hg38_chr22_bed_gz

Type (set all): Genome (set all):

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

The screenshot shows the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow' (highlighted), 'Shared Data', 'Visualization', 'Help', 'User', and 'Using 4.4 MB'. Below the navigation bar, the 'Your workflows' section contains two buttons: 'Create new workflow' and 'Upload or import workflow'. A table with the following structure is visible:

Name	# of Steps
Genes-Repeats	9

A context menu is open over the 'Genes-Repeats' workflow name, listing the following actions: Edit, Run (highlighted in orange), Share or Publish, Download or Export, Copy, Rename, View, and Delete.

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_

Step 2: Input dataset
BED file defining repetitive regions.

Repeats

3: http://bit.ly/SimpleRepeats_ucsc_hg38_chr22_bed_

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_

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

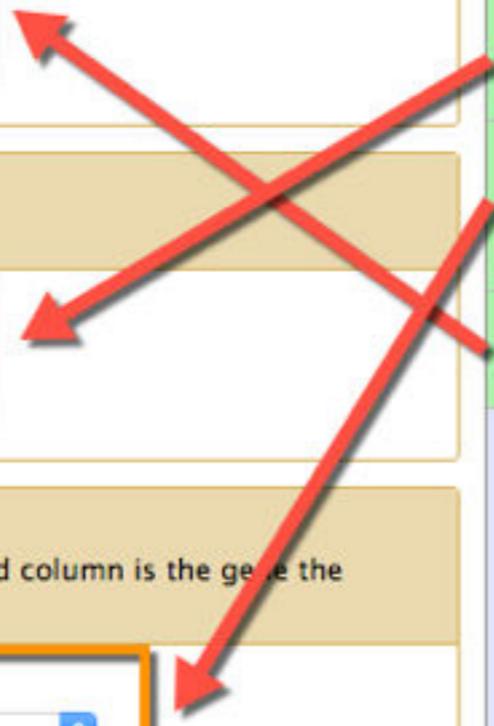
3 shown

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

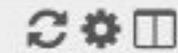
All tasks are queued



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
ASCC2	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/transcript> GeneSymbol ucsc hg38 chr22 tabular

1: <http://bit.ly/transcript>

And finally give the output dataset and the history meaningful names

The screenshot shows a 'History' panel in a software application. At the top, there is a search bar labeled 'search datasets' with a refresh icon, a settings gear, and a window icon. Below the search bar, the main history is displayed as a list of steps. The first step is highlighted in light blue and shows the dataset 'Genes-SimpleRepeats Overlaps, chr22' with a size of 1.61 MB and icons for selection, deletion, and chat. The subsequent steps (2-9) are highlighted in light green and represent a workflow of data processing operations, each with icons for visibility, editing, and deletion.

Step	Description	Visibility	Edit	Delete
	Genes-SimpleRepeats Overlaps, chr22 9 shown 1.61 MB	<input checked="" type="checkbox"/>		
9	<u>Genes w max # of overlapping SimpleRepeats, chr22</u>	<input type="checkbox"/>		<input type="checkbox"/>
8	<u>Group on data 7</u>	<input type="checkbox"/>		<input type="checkbox"/>
7	<u>Join two Datasets on data 2 and data 6</u>	<input type="checkbox"/>		<input type="checkbox"/>
6	<u>Group on data 5</u>	<input type="checkbox"/>		<input type="checkbox"/>
5	<u>Join on data 3 and data 4</u>	<input type="checkbox"/>		<input type="checkbox"/>
4	<u>Gene BED To Exon/Intron/Codon BED on data 1</u>	<input type="checkbox"/>		<input type="checkbox"/>
3	<u>http://bit.ly/SimpleRepeats ucsc hg38 chr22 bed</u>	<input type="checkbox"/>		<input type="checkbox"/>
2	<u>http://bit.ly/transcript GeneSymbol ucsc hg38 chr22 tabul</u>	<input type="checkbox"/>		<input type="checkbox"/>

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

Galaxy on Jetstream

NEW



Instances

<input checked="" type="checkbox"/>	Name	Status	Activity	IP Address	Size	Provider
<input checked="" type="checkbox"/>	 Galaxy 16.01 Standalone	● Active		129.114.17.100	M1.Large	Jetstream - TACC

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



Join us in beautiful

Bloomington, Indiana

for the 2016 Galaxy
Community Conference
and pre-conference activities!

June 25-29, 2016



Considered one of the five
prettiest campuses in the US,
Indiana University is one of
the major public research
universities in the nation, and
home to the National Center
for Genome Analysis Support.



galaxyproject.org/gcc2016



26 - 30 June France

GCC 2017 Montpellier



Le Corum
Conference centre

gcc2017.sciencesconf.org

November 7-11



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

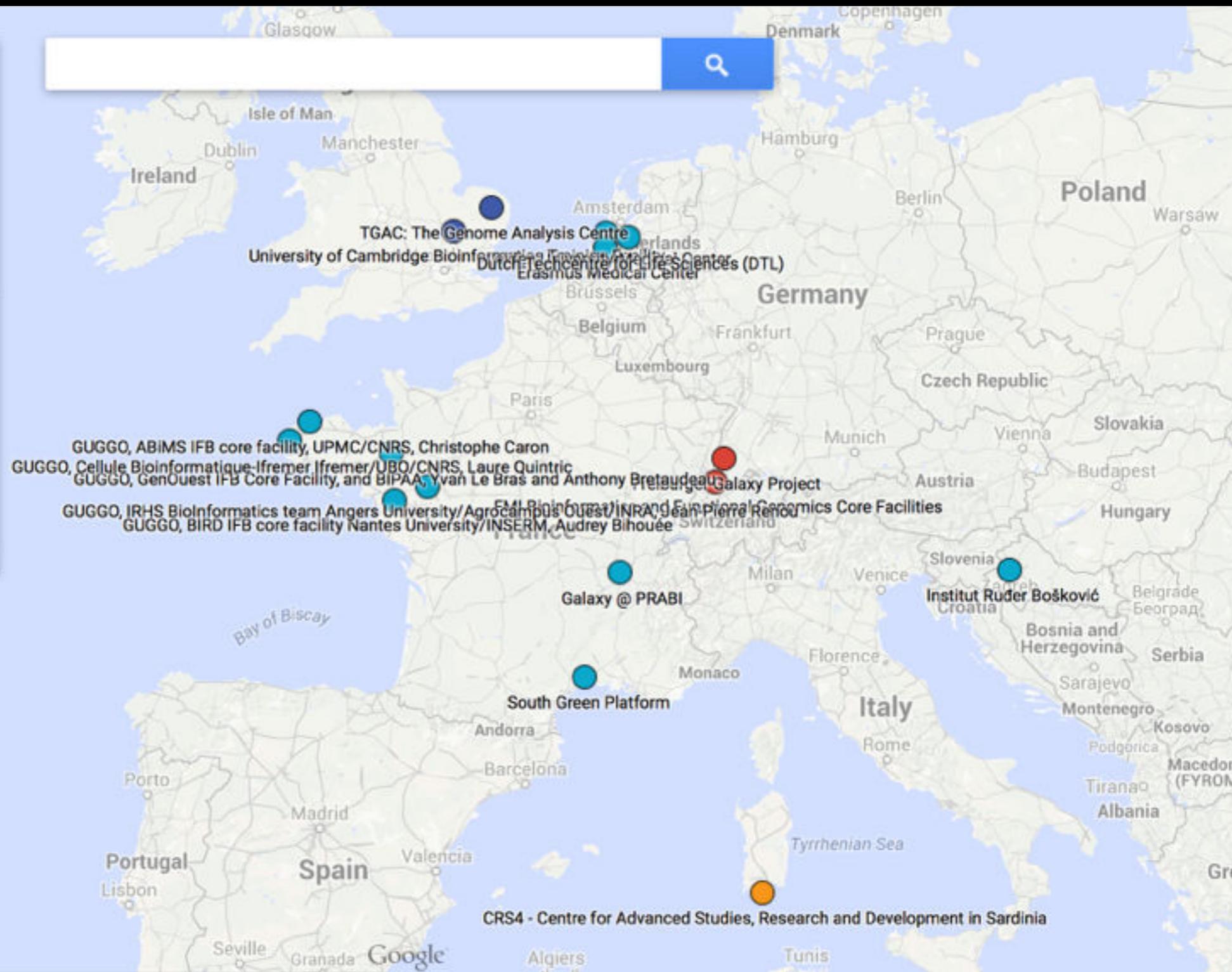
Galaxy Training Network: Trainer Locations

The Galaxy Training Network
(<https://wiki.galaxyproject.org/Teach/GTN>)

Made with Google My Maps

Trainers

- Global
- Regional
- Local
- Continental
- Institution



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