

Introduction to Galaxy

Boyce Thompson Institute & Cornell University
July 21, 2016

Dave Clements

Galaxy Team

Johns Hopkins University

<http://galaxyproject.org/>



#usegalaxy @galaxyproject



Cornell University



Agenda

- 9:00 Welcome
- 9:20 Basic Analysis with Galaxy
- 10:45 Break
- 11:00 Basic Analysis into Reusable Workflows
- 12:20 Lunch (on your own)
- 1:20 RNA-Seq Analysis, Part I
- 2:50 Break
- 3:05 RNA-Seq Analysis, Part II
- 4:30 Launch your own Galaxy with AWS
- 5:00 Done

bit.ly/btigxy

Goals

Provide an introduction to using Galaxy for bioinformatic analysis. Demonstrate how Galaxy can help you explore and learn options, perform analysis, and then share, repeat, and reproduce your analyses.

This workshop does cover RNA-Seq but you won't be an expert at the end of the workshop. You will know enough to get started, and how to use Galaxy to learn more.

What is Galaxy?

Keith Bradnam's definition:

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

From

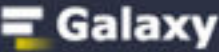
"13 Questions You May Have About Galaxy"

<http://bit.ly/13questions>

Galaxy is available several ways ...

<http://galaxyproject.org>

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



Analyze DataWorkflowShared DataVisualizationHelpUser

Using 0%

Galaxy now runs some (larger, multicore) jobs on [Jetstream](#), you may encounter a few problems related to this. We are working on these, and please feel free to report any errors you encounter.


Tools


search tools

[Get 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](#).

Want help?
Get answers.

 **Biostars**
GALAXY EXPLAINED




Tweets by @galaxyproject

**Galaxy Project** @galaxyproject


Tutorial & history from today's #TAGC16 #usegalaxy workshop are at:
bit.ly/TAGC_GXY_PDF
[twitter.com/galaxyproject/...](https://twitter.com/galaxyproject/)


16 Jul


**Galaxy Project** @galaxyproject


omorrow morning 8am #TAGC16: An Introduction to Using Galaxy for Genetic Data Analysis [genetics-gsa.org/genetics/2016/...](http://genetics-gsa.org/genetics/2016/)

[Embed](#)[View on Twitter](#)

 PENNSTATE

 JOHNS HOPKINS

 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

Galaxy / Metabiome Portal



The Microbiome Analysis Center
Life on a Smaller Scale

Welcome to the Metabiome Portal @ GMU

We have developed the MMC Metabiome Portal, a flexible and extensible web browser, with the ability to simplify, control, integrate, compare, and analyze all microbiome and metagenomic data. The portal is a unified database management system and data-based analytical resources and includes several tools such as: taxonomic clustering,...

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

BGI

(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



We are using this today

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

<http://globus.org/>

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

Galaxy on the Cloud: Galaxy CloudMan

<http://usegalaxy.org/cloud>

- Start with a **fully configured and populated** (tools and data) Galaxy instance.
- Allows you to scale up and down your compute assets as needed.
- Someone else manages the data center



CLOUDMAN

Agenda

9:00 Welcome

9:20 Basic Analysis with Galaxy

10:45 Break

11:00 Basic Analysis into Reusable Workflows

12:20 Lunch (on your own)

1:20 RNA-Seq Analysis, Part I

2:50 Break

3:05 RNA-Seq Analysis, Part II

4:30 Launch your own Galaxy with AWS

5:00 Done

bit.ly/btigxy

Quick Poll: Are you ...

1. A bioinformatics novice

2. A bioinformatics apprentice

3. A bioinformatics guru

Yes, those are your only choices.

<http://galaxyproject.org>

Basic Analysis

Which exons have most overlapping
Repeats?

Use Human, HG38, GENCODE v24,
Chromosome 22

cloud1.galaxyproject.org

cloud2.galaxyproject.org

cloud5.galaxyproject.org

Exons & Repeats: A General Plan

- Get some data
 - Get Data → UCSC Table Browser
- Identify which exons have Repeats
- Count Repeats per exon
- Visualize, save, download, ... exons with most Repeats

(~ <http://usegalaxy.org/galaxy101>)



Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersection of DNA sequence covered by a track. For help in using this application see [Using the Table Browser](#) this form, the [User's Guide](#) for general information and sample queries, and the OpenHelix Table Browser presentation of the software features and usage. For more complex queries, you may want to use the [Query Builder](#). To examine the biological function of your set through annotation enrichments, send the data to [BioMart](#) for use with diverse computational tools. Refer to the [Credits](#) page for the list of contributors and the [FAQ](#) for these data. All tables can be downloaded in their entirety from the [Sequence and Annotation Download](#) page.

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

group: **track:**

table:

region: ☐ genome ☒ position

identifiers (names/accessions):

filter:

intersection:

correlation:

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

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

file type returned: ☒ plain text ☐ gzip compressed



Output knownGene as BED

☐ Include [custom track](#) header:

name=

description=

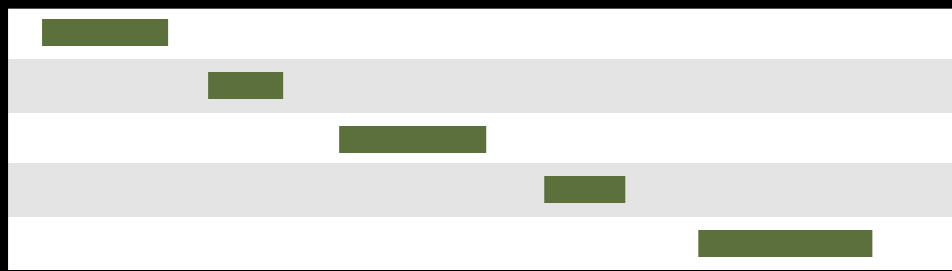
visibility=

url=

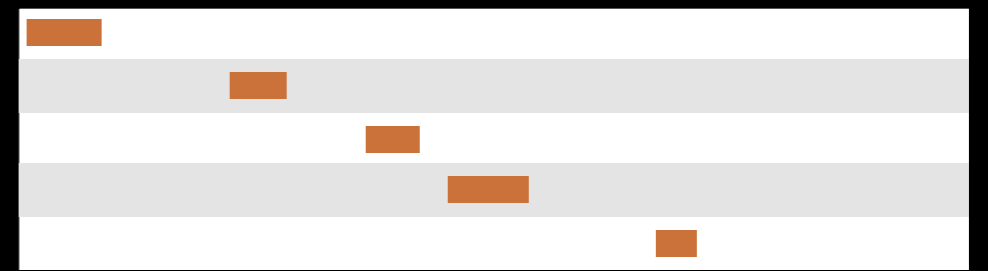
Create one BED record per:

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

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream in order to avoid extending past the edge of the chromosome.

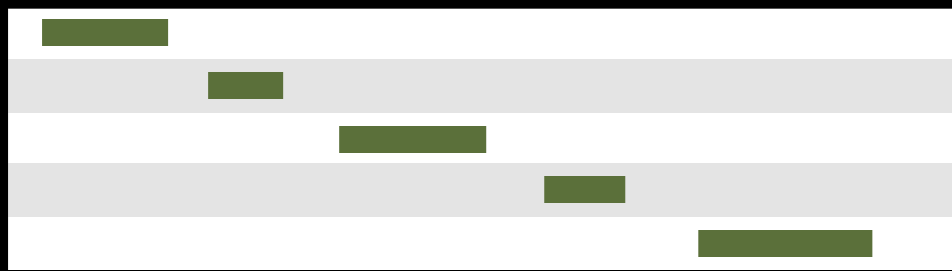


Exons

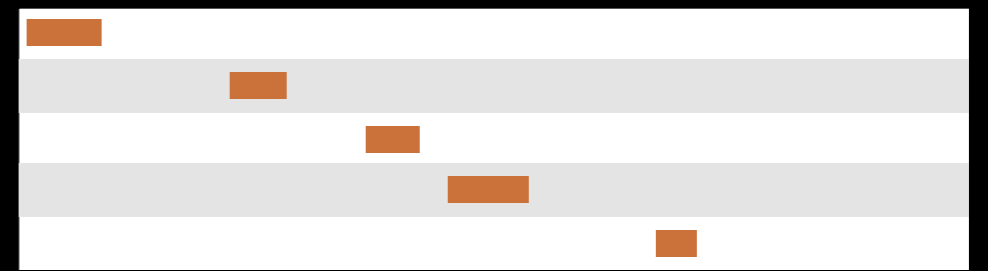


Repeats

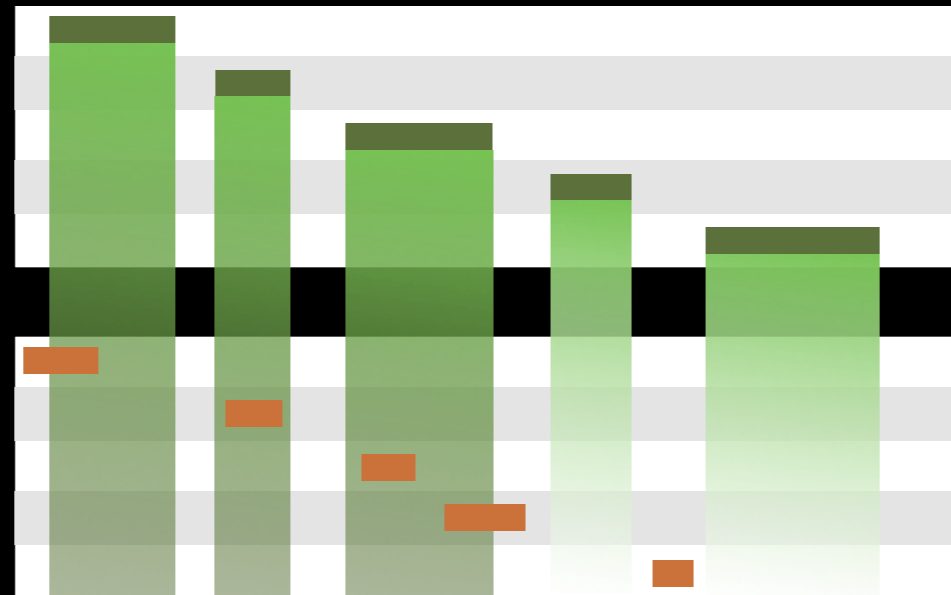
(Identify which exons have Repeats)



Exons



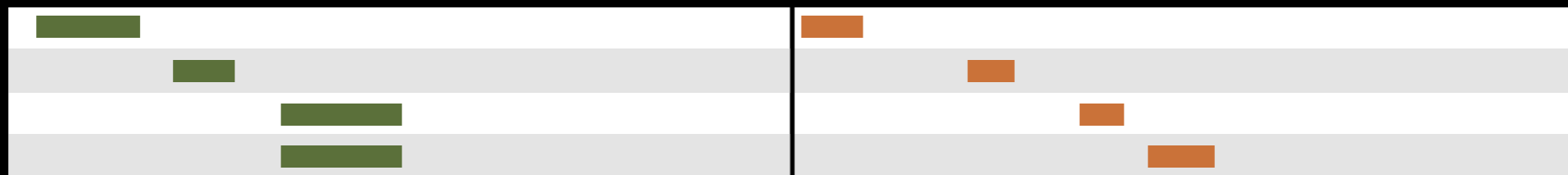
Repeats



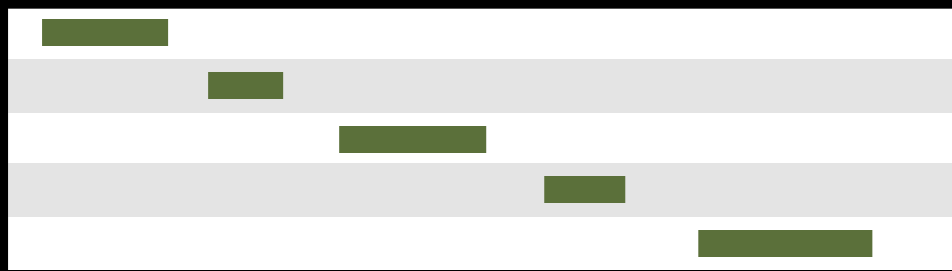
Exons

Repeats

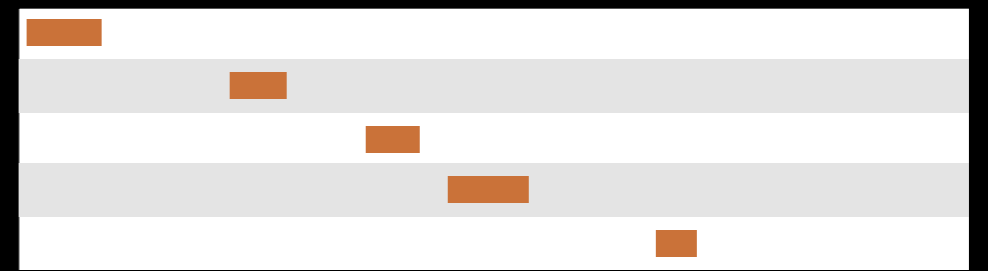
Overlap pairings



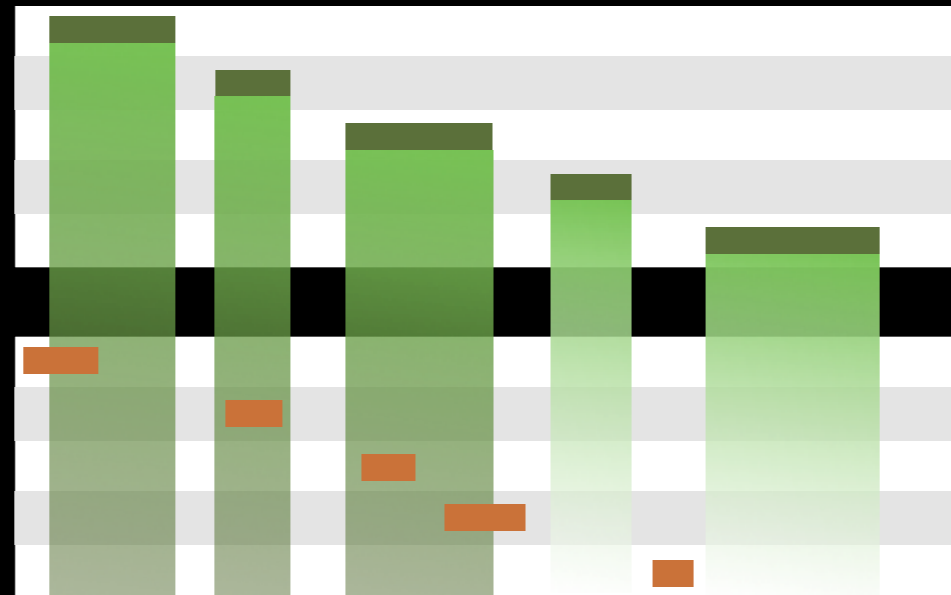
Operate on Genomic Intervals → Join
(Identify which exons have Repeats)



Exons



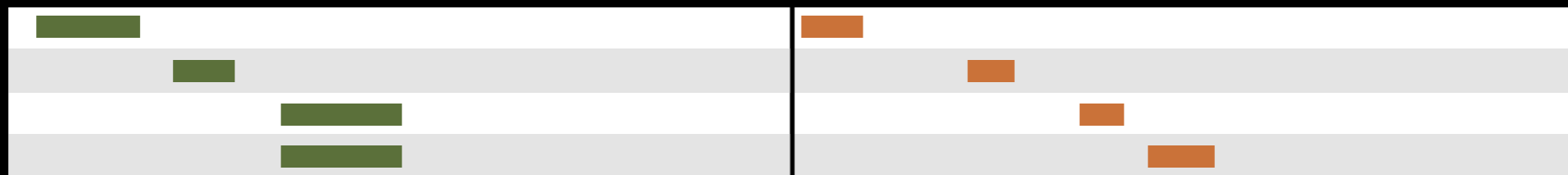
Repeats



Exons

Repeats

Overlap pairings



(Count Repeats per exon)



Exon overlap counts

Join, Subtract, and Group → Group

Published History: Exons with overlapping repeats, basic

Yay!

We have exon names and counts!

We are now going to extend that work.

Let's **create a copy** of this history that we will extend.

Exons & Repeats: Pick an Exercise

1. Report the number of overlapping repeats **every exon** has (including exons with **0** overlapping repeats.)
2. Output the **list of exons** that have overlapping repeats, **in BED format**. Set the score column be the number of overlapping repeats that exon has.

Everything you need will be in these toolboxes

- Text manipulation (cut is particularly useful)
- Operate on genomic intervals
- Join, subtract and group
- Filter and sort

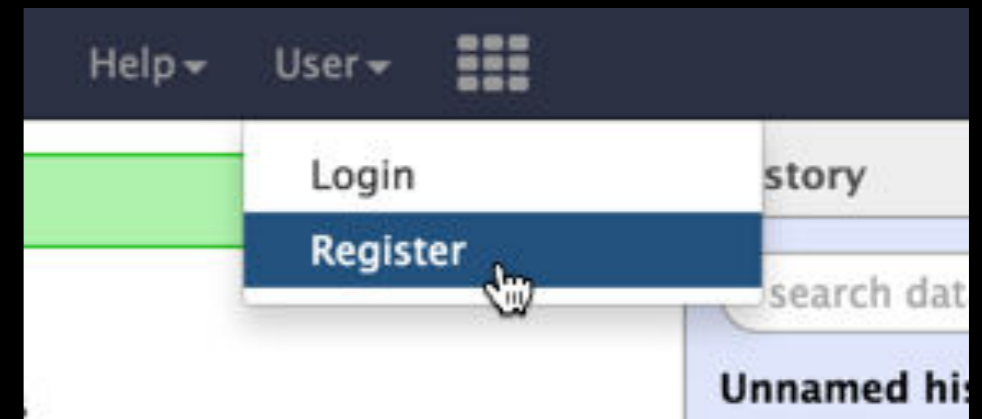
But first, create a login

Don't need to login to use Galaxy, but do need one to use all its features

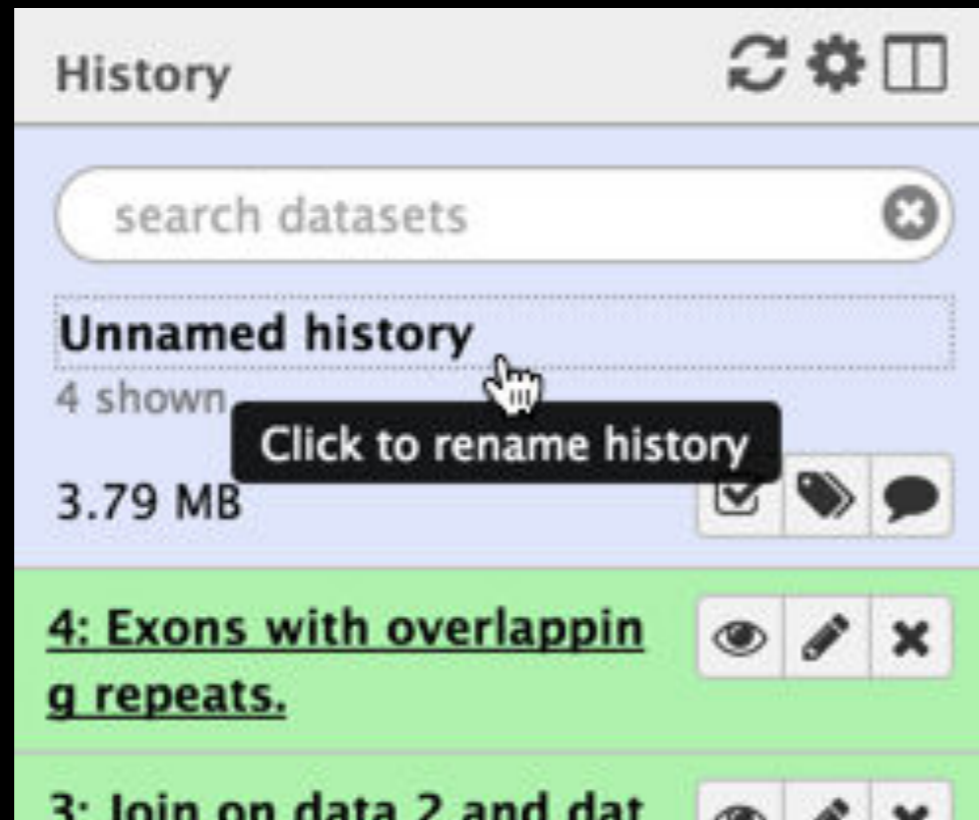
Use an email address you can remember.

Use a low security password.

This account will go away on Wednesday night.

A screenshot of a 'Create account' form. The form has a yellow header bar with the title 'Create account'. It contains four input fields: 'Email address:', 'Password:', 'Confirm password:', and 'Public name:'. Below the 'Public name' field is a text explanation: 'Your public name is an identifier that will be used to generate a URL for you to share publicly. Public names must be at least three characters long and can only contain lower-case letters, numbers, and the '-' character.' A 'Submit' button is located at the bottom left of the form.

Second, name your existing history



Give your existing history a meaningful name.

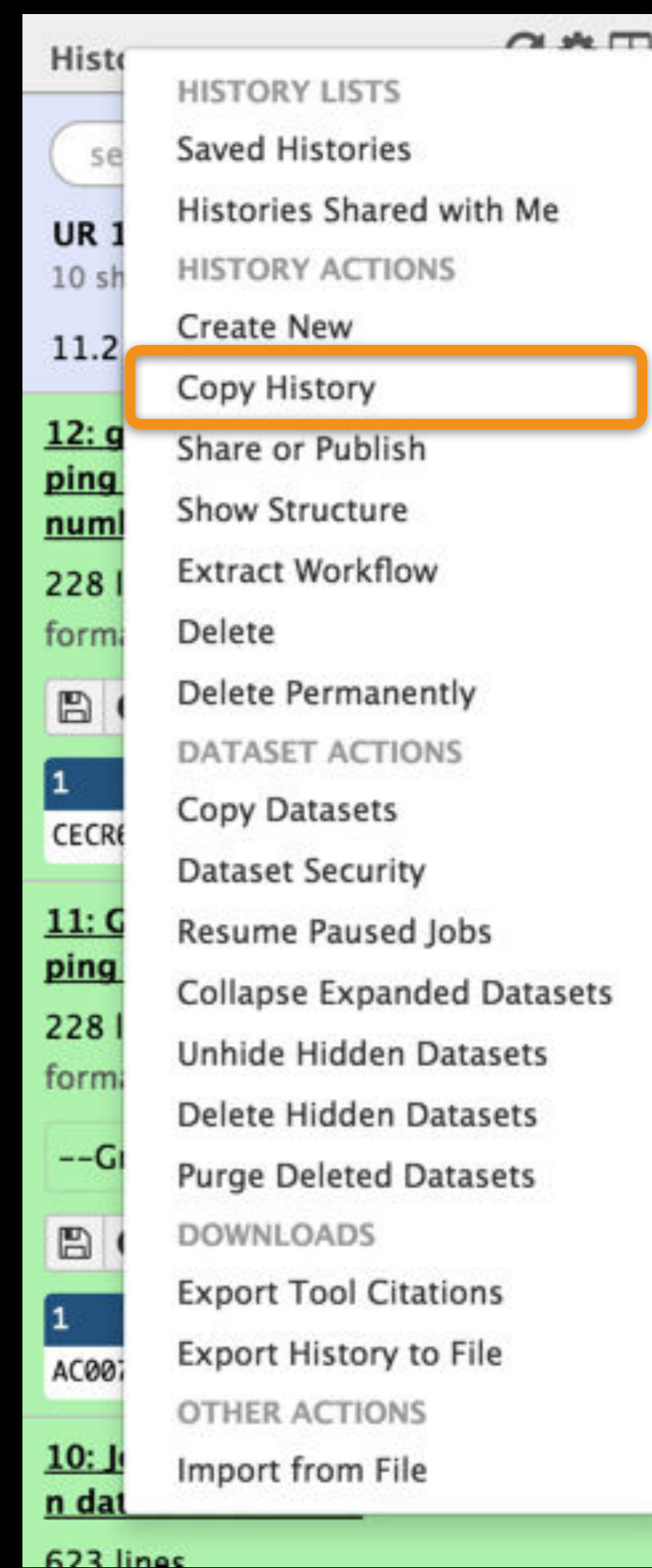
3rd, make a copy of your history



(cog) → Copy History

Name the copy based on the exercise you pick

Becomes your new current history.



All exons, even those with no overlap

Can take advantage of fact that scores are already 0.
Join, subtract and group not a bad place to start.

Published History: Exons with number of overlapping
repeats, including 0

List of exons with overlaps, in BED

Can be done in two steps, one of them a Cut,
plus an edit attributes step at the end:

The screenshot displays the Galaxy web interface. On the left, the 'Datatype' tab is selected and highlighted with an orange box. It shows a 'Change data type' section with a 'New Type:' dropdown menu set to 'bed', also highlighted with an orange box. Below the dropdown, a message states: 'This will change the datatype of the existing dataset but *not* modify its contents. Use this if Galaxy has incorrectly guessed the type of your dataset.' A 'Save' button is at the bottom left. On the right, the 'History' panel shows a list of datasets. The top entry is 'Exons with overlapping repeats, in BED' (3.92 MB). Below it, a green entry is titled '6: Exons with overlapping repeats, in BED' (792 regions, format: interval, database: hg38). This entry has an edit icon (pencil) highlighted with an orange box. At the bottom of the history panel, a table header is visible: '1. Chrom 2. Start 3. End 4. Name'.

Published History: Exons with overlapping repeats, in BED

Agenda

9:00 Welcome

9:20 Basic Analysis with Galaxy

10:45 Break

11:00 Basic Analysis into Reusable Workflows

12:20 Lunch (on your own)

1:20 RNA-Seq Analysis, Part I

2:50 Break

3:05 RNA-Seq Analysis, Part II

4:30 Launch your own Galaxy with AWS

5:00 Done



gmod.org

bit.ly/btigxy

Agenda

- 9:00 Welcome
- 9:20 Basic Analysis with Galaxy
- 10:45 Break
- 11:00 Basic Analysis into Reusable Workflows**
- 12:20 Lunch (on your own)
- 1:20 RNA-Seq Analysis, Part I
- 2:50 Break
- 3:05 RNA-Seq Analysis, Part II
- 4:30 Launch your own Galaxy with AWS
- 5:00 Done

bit.ly/btigxy

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:

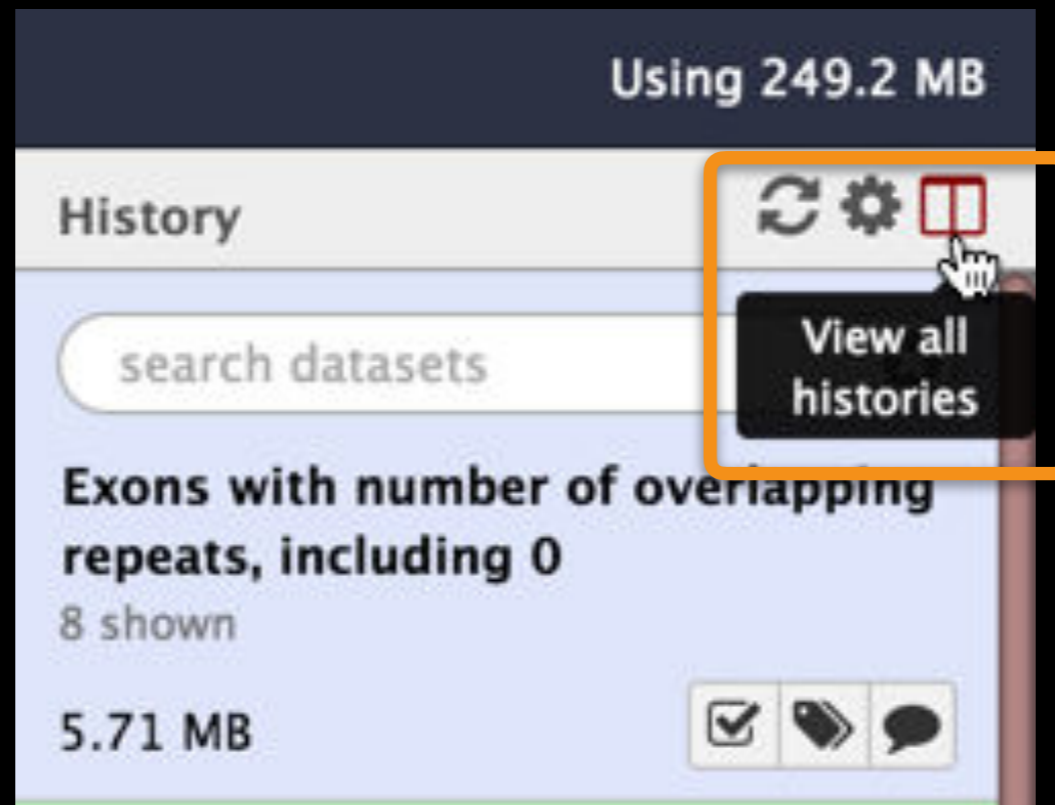
A series of analysis steps

Can be repeated with different data

Exons and Repeats *History* → Reusable *Workflow*?

- The analysis we just finished was about
 - Human chr22
 - Overlap between exons and repeats
 - And then rolling that up to genes
- But, ...
 - is there anything inherent in the analysis **about humans, exons or repeats?**

Get back to the original history



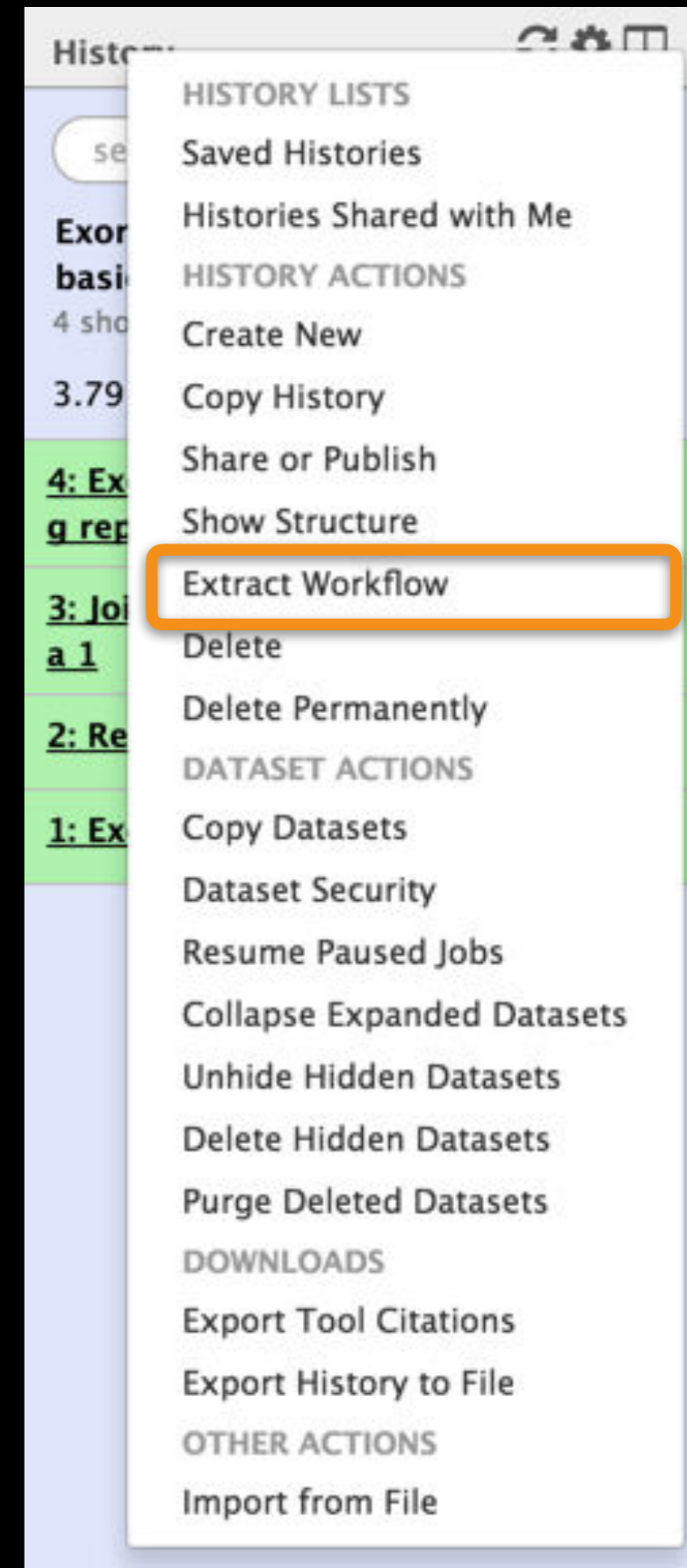
Create a Workflow from a History

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: ...

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

Workflow constructed from history 'Exons with overlapping repeats, basic'

Create Workflow

Check all

Uncheck all

Tool

History items created

UCSC Main

This tool cannot be used in workflows



1: Exons, chr22

☒ Treat as input dataset

UCSC Main

This tool cannot be used in workflows



2: Repeats, chr22

☒ Treat as input dataset

Join

☒ Include "Join" in workflow



3: Join on data 2 and data 1

Group

☒ Include "Group" in workflow



4: Exons with overlapping repeat
s.

Workflow editor

Tools

search tools

Inputs

[Get Data](#)

[Send Data](#)

[Lift-Over](#)

[Text Manipulation](#)

[Filter and Sort](#)

[NGS: QC and manipulation](#)

[NGS: DeepTools](#)

[NGS: Mapping](#)

[NGS: RNA Analysis](#)

[NGS: SAM Tools](#)

[NGS: BAM Tools](#)

[NGS: Picard](#)

[NGS: Variant Analysis](#)

[NGS: VCF Manipulation](#)

[NGS: ChIP-seq](#)

[Join, Subtract and Group](#)

[Operate on Genomic Intervals](#)

[BEDtools](#)

[Convert Formats](#)

[FASTA manipulation](#)

[Extract Features](#)

[Fetch Sequences](#)

[Fetch Alignments](#)

Workflow Canvas | count overlapping features

```
graph LR; A[Input dataset output] --> C[Join with output interval]; B[Input dataset output] --> C; C --> D[Group Select data out_file1 tabular];
```

Details

Edit Workflow Attributes

Name:
count overlapping features

Tags:

Apply tags to make it easy to search for and find items with the same tag.

Annotation / Notes:
Describe or add notes to workflow
Add an annotation or notes to a workflow; annotations are available when a workflow is viewed.

Published Workflow: Feature Overlap Counting

Workflow editor: save your changes

The screenshot displays a workflow editor interface. On the left is a 'Tools' sidebar with a search bar and a list of tool categories: Inputs, Get Data, Send Data, Lift-Over, Text Manipulation, Filter and Sort, NGS: QC and manipulation, NGS: DeepTools, NGS: Mapping, NGS: RNA Analysis, NGS: SAM Tools, NGS: BAM Tools, NGS: Picard, NGS: Variant Analysis, NGS: VCF Manipulation, NGS: ChIP-seq, Join, Subtract and Group, Operate on Genomic Intervals, BEDtools, Convert Formats, FASTA manipulation, Extract Features, Fetch Sequences, and Fetch Alignments. The main area is the 'Workflow Canvas | count overlapping features', which has a grid background. It contains two 'Input dataset' tools, each with an 'output' field, connected by lines to a central 'Join' tool. The 'Join' tool has a 'with' field and an 'output (interval)' field. A context menu is open over the 'Join' tool, showing options: Save, Run, Edit Attributes, Auto Re-layout, and Close. The 'Details' panel on the right shows 'Edit Workflow Attributes' and 'Workflow Attributes' sections. A small preview window at the bottom right shows a grid of colored squares.

Published Workflow: Feature Overlap Counting

Workflow Testing

Guided: rerun with same inputs

Workflow → Run

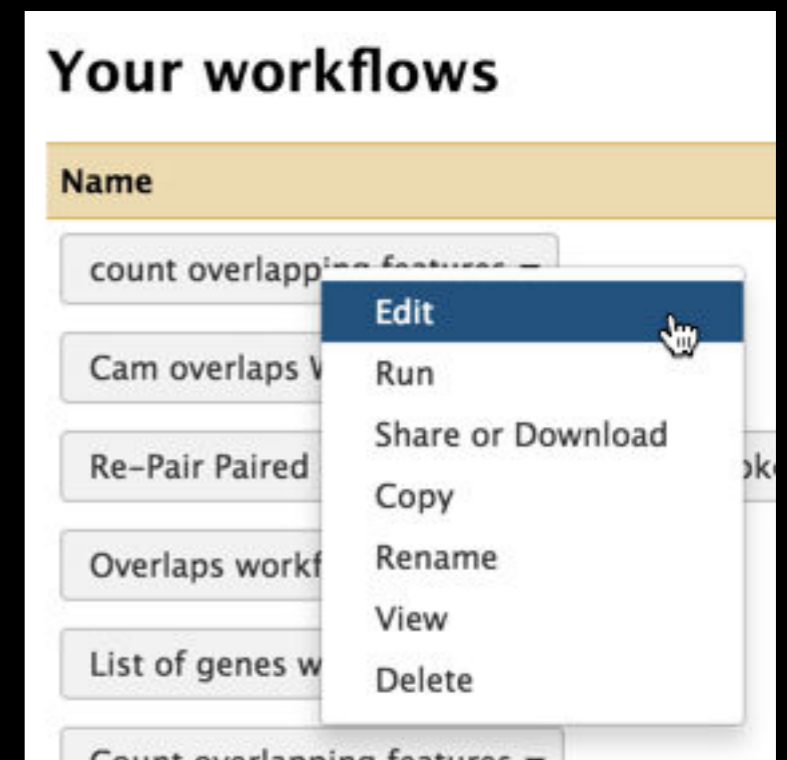
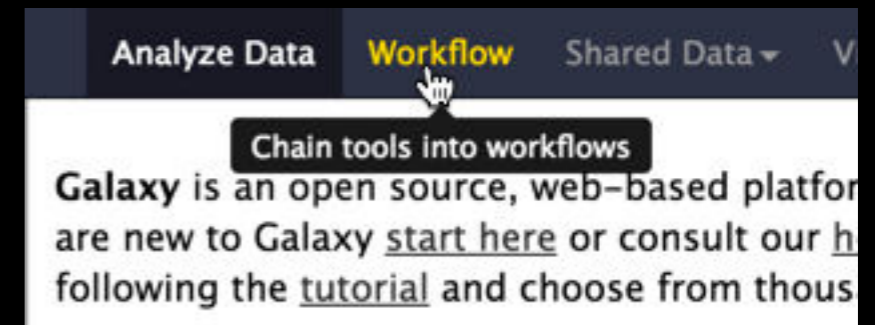
Did that work?

On your own:

Count # of exons overlapping each repeat

Did that work? *Why not?*

Edit workflow: doc assumptions



Published Workflow: Feature Overlap Counting

Workflows: Sweet spots

Short, well-defined tasks, with well-defined inputs and outputs.

Analysis pipelines for large experiments with many samples where sample and data preparation protocols are the same throughout.

Agenda

- 9:00 Welcome
- 9:20 Basic Analysis with Galaxy
- 10:45 Break
- 11:00 Basic Analysis into Reusable Workflows
- 12:20 Lunch (on your own)**
- 1:20 RNA-Seq Analysis, Part I
- 2:50 Break
- 3:05 RNA-Seq Analysis, Part II
- 4:30 Launch your own Galaxy with AWS
- 5:00 Done



bit.ly/btigxy

Agenda

- 9:00 Welcome
- 9:20 Basic Analysis with Galaxy
- 10:45 Break
- 11:00 Basic Analysis into Reusable Workflows
- 12:20 Lunch (on your own)
- 1:20 RNA-Seq Analysis, Part I**
- 2:50 Break
- 3:05 RNA-Seq Analysis, Part II
- 4:30 Launch your own Galaxy with AWS
- 5:00 Done

bit.ly/btigxy

Quick Poll: Are you ...

1. An RNA-Seq novice
2. An RNA-Seq apprentice
3. An RNA-Seq guru

Yes, those are your only choices.

<http://galaxyproject.org>

RNA-Seq Analysis: Get the Data

Shared Data → Data Libraries → Training → RNA-Seq*

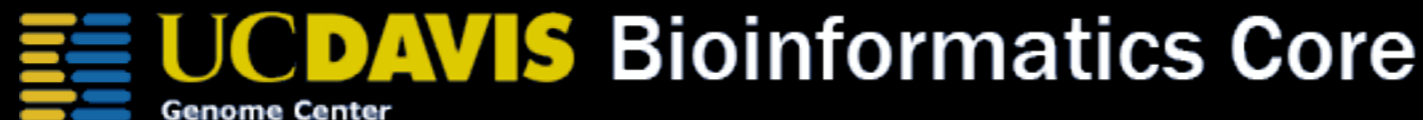
→ UC-Davis → Raw Reads

Select first two

MeOH_REP1_R1

MeOH_REP1_R2

Import into a new history



* RNA-Seq example datasets from the 2013 UC Davis Bioinformatics Short Course. <http://bit.ly/ucdbsc2013>

NGS Data Quality Control

- FASTQ format
- Examine quality in an RNA-Seq dataset
- Trim/filter as we see fit, hopefully without breaking anything.

Quality Control is not sexy.

But it is vital.

What is FASTQ?

- Specifies sequence (FASTA) and quality scores (PHRED)
- Text format, 4 lines per entry

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
! ' ' * ( ( ( ( * * * + ) ) % % % + + ) ( % % % % ) . 1 * * * - + * ' ' ) ) * * 55CCF>>>>>CCCCCCC65
```

- **FASTQ is such a cool standard, there are 3 (or 5) of them!**

[illegible]

http://en.wikipedia.org/wiki/FASTQ_format

NGS Data Quality: Assessment tools

NGS QC and Manipulation → **FastQC**

Generates summary quality information.

FastQC Read Quality reports (Galaxy Version 0.63)

VersionsOptions

Short read data from your current history

1: MeOH_REP1_R1.fastq

Contaminant list

Nothing selected

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Submodule and Limit specifying file

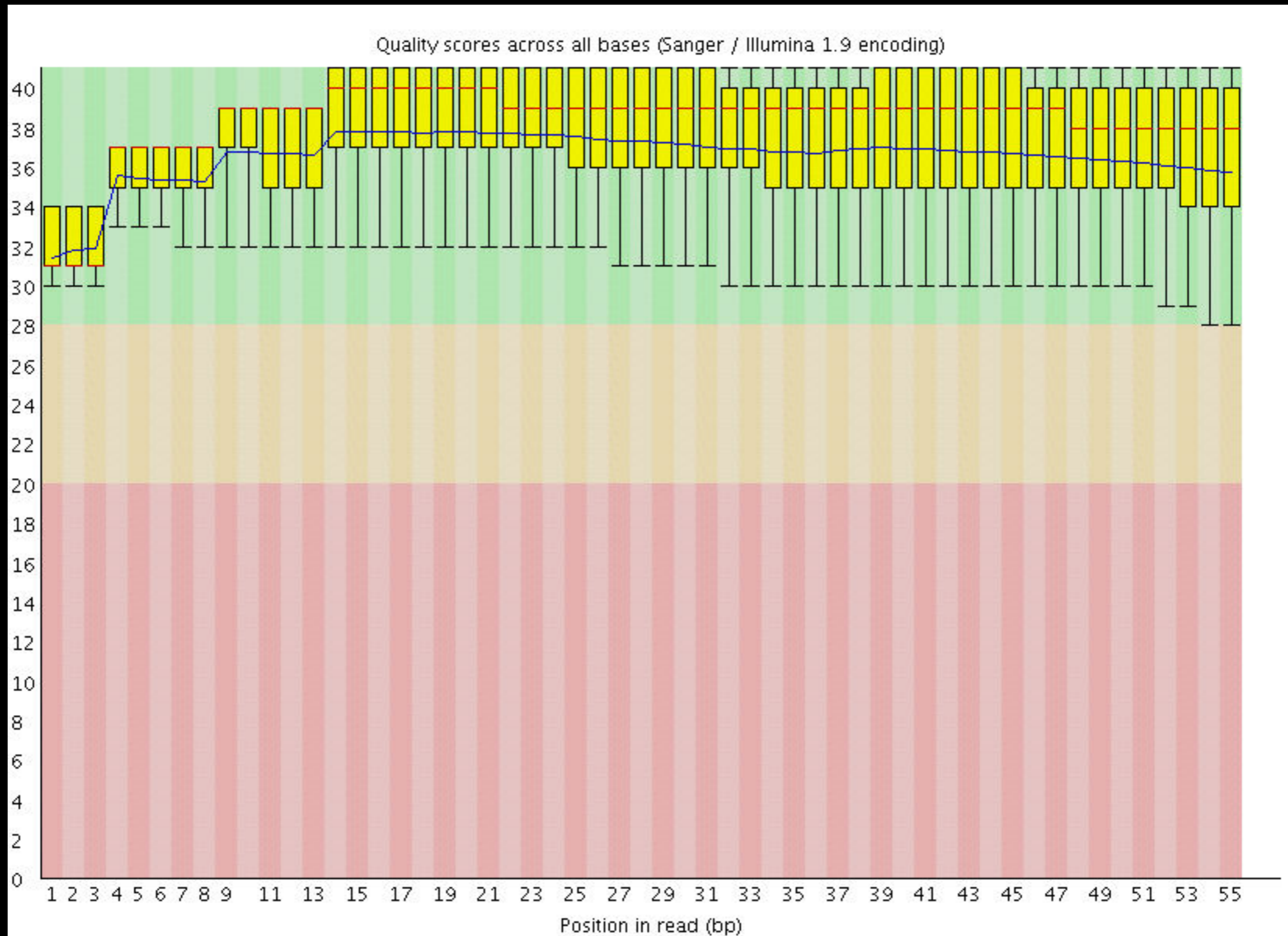
Nothing selected

a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

✓ Execute

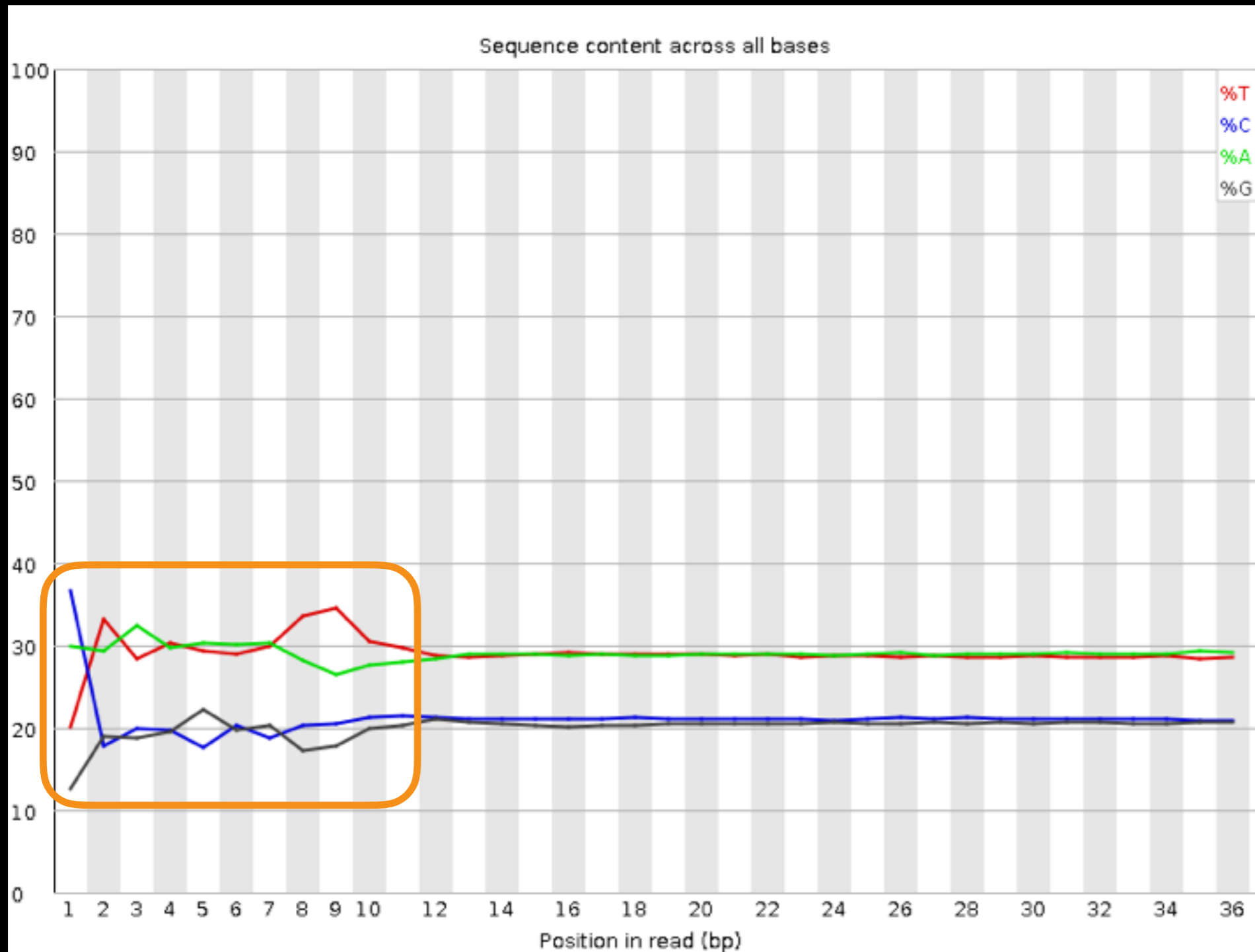
<http://bit.ly/FastQCBoxPlot>

NGS Data Quality: Assessment tools



<http://bit.ly/FastQCBoxPlot>

NGS Data Quality: Sequence bias at front of reads?



From a sequence specific bias that is caused by use of random hexamers in library preparation.

Hansen, *et al.*, "Biases in Illumina transcriptome sequencing caused by random hexamer priming" *Nucleic Acids Research*, Volume 38, Issue 12 (2010)

Common Trimming options

- **Drop the first n columns** from your reads
- **Drop the last n columns** from your reads
- **Sliding window** approach: only keep regions that are above a specified quality threshold
- **Keep or drop whole read** based on overall quality

Common Trimming Pitfalls

Broken Pairs

Often, one side of a pair passes QC, while the other does not.

Broken pairings can affect results in subtle or drastic ways

Short short reads.

QC may reduce reads to a length at which their mapping is no longer meaningful.

Need help with Trimming? (and anything else)

That's a **whole lotta options...**

Choices you make now have impact on downstream tools

NGS = a whole lotta options in general

What to do?

How to better understand bioinformatics & Galaxy

- **Experiment.** (You are already used to the idea and)
Galaxy makes it easy
- **Read** tool documentation and tool and method review papers
- **Get Help!**
 - <http://biostars.org/>
 - <http://seqanswers.com/>
 - <https://biostar.usegalaxy.org/>
 - <http://galaxyproject.org/search>



Trimmomatic to the rescue

Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Tool Version 0.32.3) Options

Paired end data?

Input Type
Pair of datasets

Input FASTQ file (R1/first of pair)
 1: MeOH_REP1_R1

Input FASTQ file (R2/second of pair)
 2: MeOH_REP1_R2

Perform initial ILLUMINACLIP step?

Cut adapter and other illumina-specific sequences from the read

Trimmomatic Operation
1: Trimmomatic Operation

Select Trimmomatic operation to perform
Sliding window trimming (SLIDINGWINDOW)

Bolger, A.M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, doi: 10.1093/bioinformatics/btu170

Trimmomatic Operation

1: Trimmomatic Operation

Select Trimmomatic operation to perform

Sliding window trimming (SLIDINGWINDOW)

Sliding window trimming (SLIDINGWINDOW)

Drop reads below a specified length (MINLEN)

Cut bases off the start of a read, if below a threshold quality (LEADING)

Cut bases off the end of a read, if below a threshold quality (TRAILING)

Cut the read to a specified length (CROP)

Cut the specified number of bases from the start of the read (HEADCROP)

Trimmomatic preserves read pairing

Multiple filters can be run in arbitrary order

We'll use **sliding window**, followed by **minimum length**.

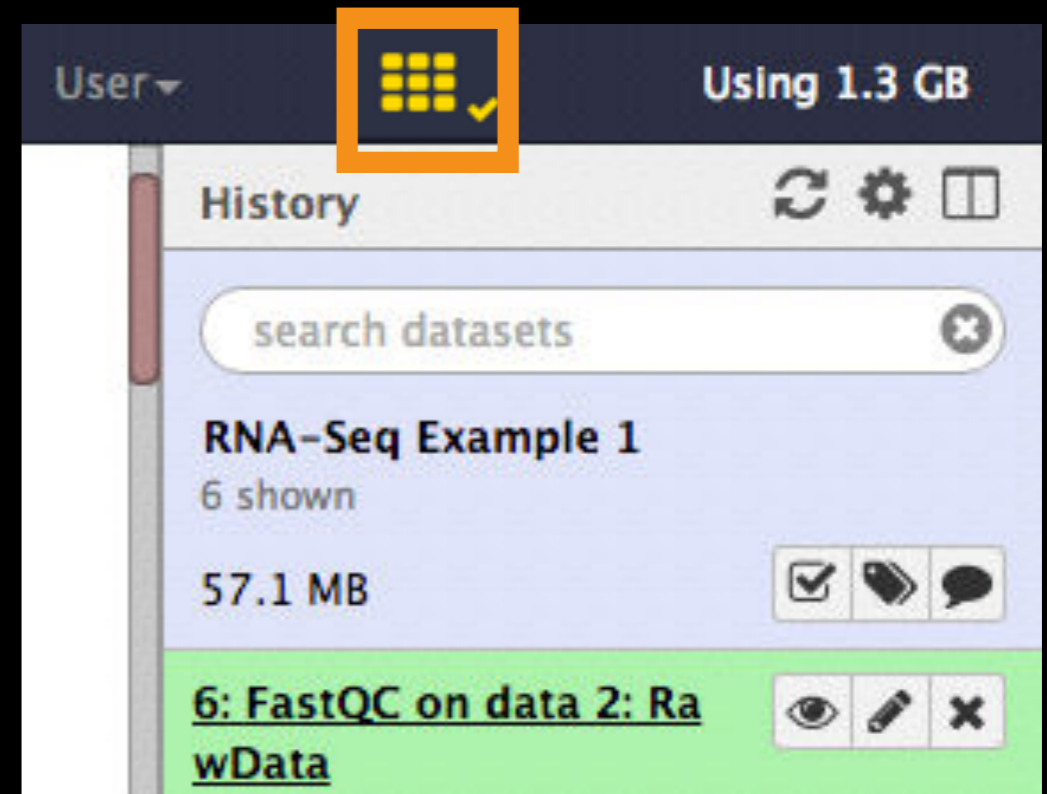
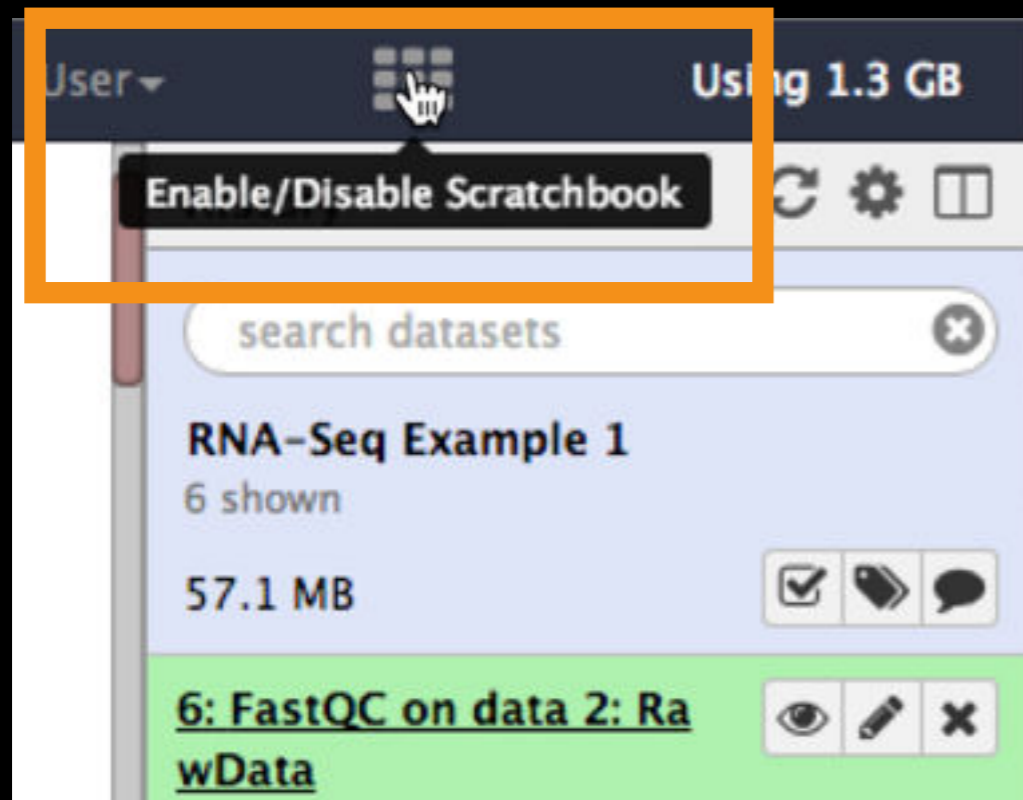
Run FastQC on post-Trimmatic Datasets

NGS QC and Manipulation → **FastQC**

Now, let's see what changed

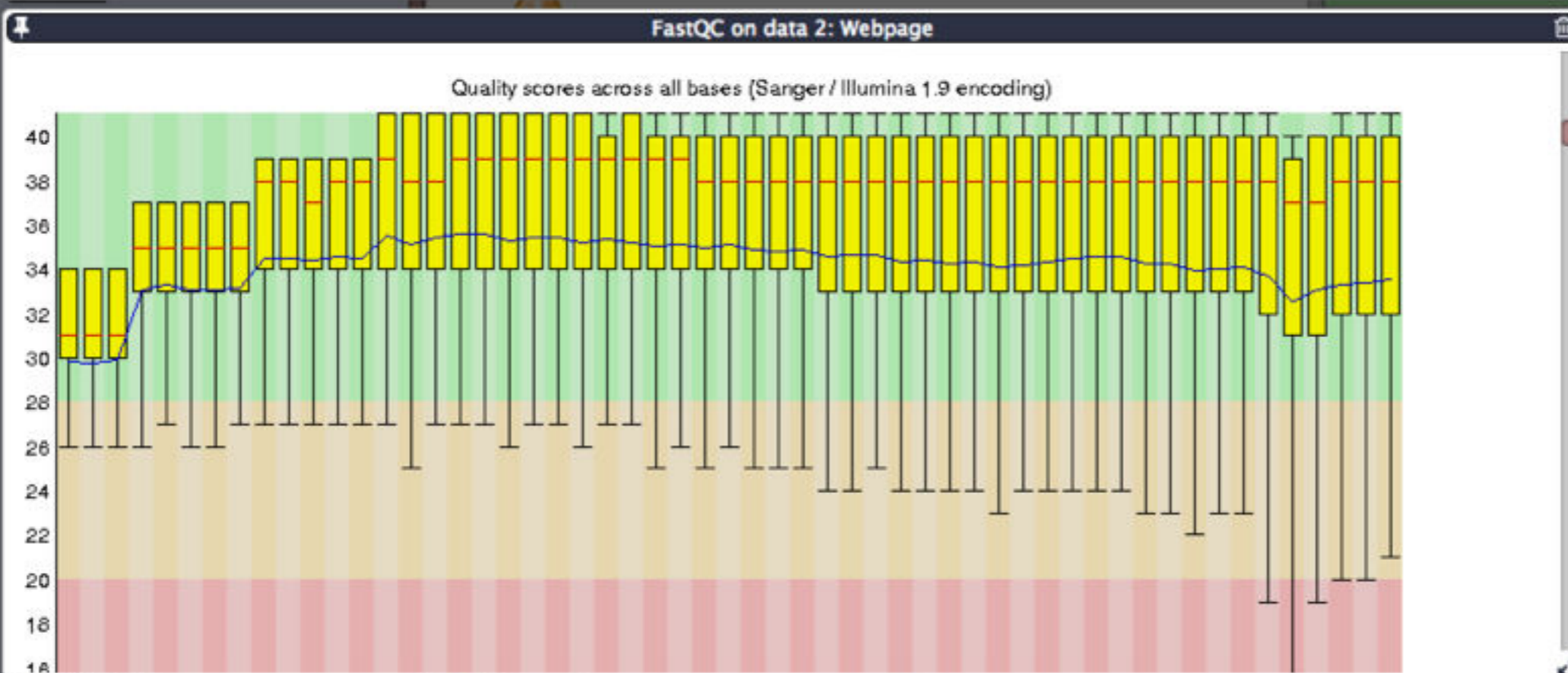
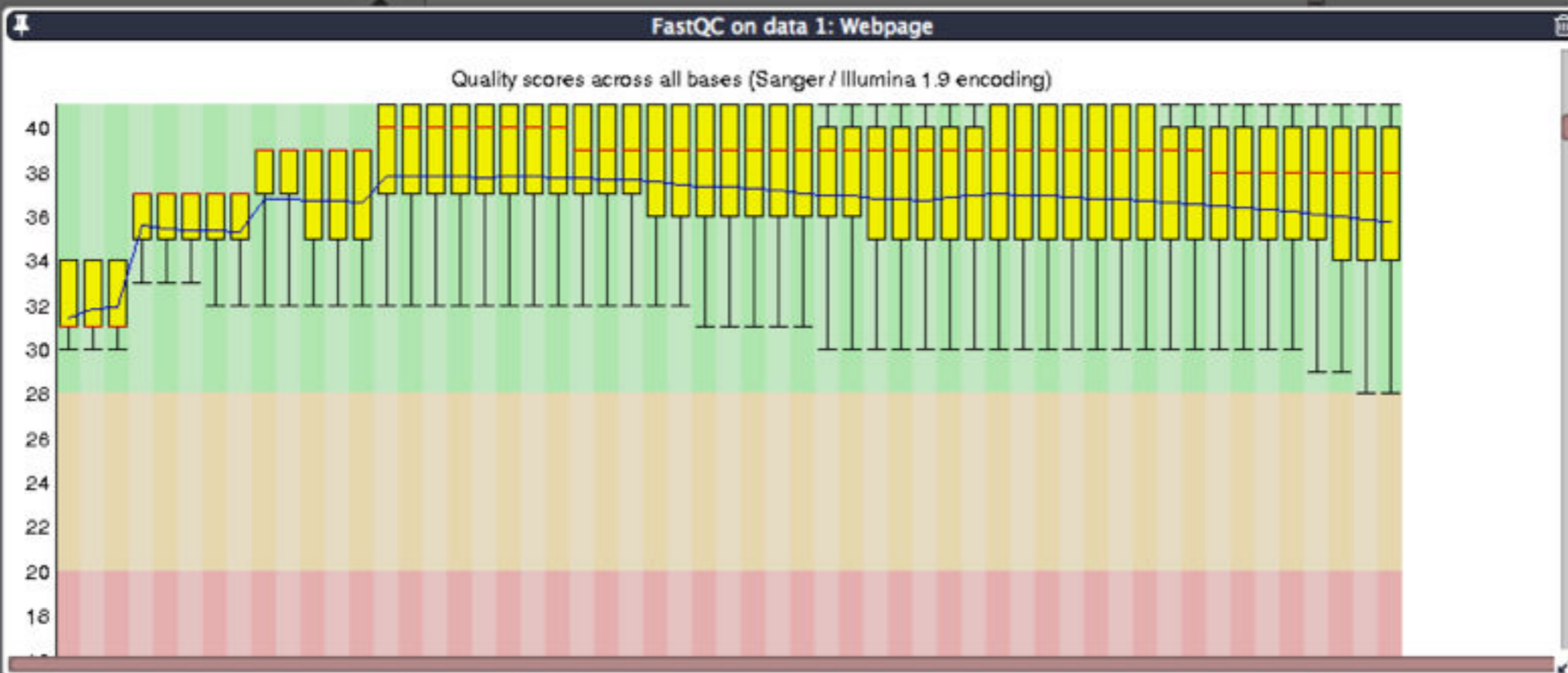
Shared History: RNA-Seq MeOH_REPI QC

Scratchbook: View multiple datasets



And the icon turns **yellow**!

Poke the **pre**-Trimmomatic reverse read FastQC report in the eye, and then poke the **post**-Trimmomatic FastQC report in the eye.



And after some resizing and scrolling you see this

NGS Data Quality Assessment

Now, just 10 more datasets to go!

Your Friend: The Multiple datasets button

Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.32.3)

Options

Paired end data?

YesNo

Input Type

Pair of datasets

Input FASTQ file (R1/first of pair)

1: MeOH_REP1_R1.fastq

Multiple datasets

(R2/second of pair)

2: MeOH_REP1_R2.fastq

Perform initial ILLUMINACLIP step?

YesNo

Cut adapter and other illumina-specific sequences from the read

Trimmomatic Operation

1: Trimmomatic Operation

Version 0.32.3)

Paired end data?

Yes

No

Input Type

Pair of datasets ▼

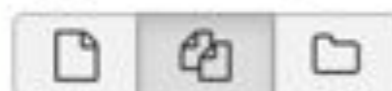
Input FASTQ file (R1/first of pair)



11: R3G_REP3_R1.fastq
10: R3G_REP2_R2.fastq
9: R3G_REP2_R1.fastq
8: R3G_REP1_R2.fastq
7: R3G_REP1_R1.fastq

This is a batch mode input field. A separate job will be triggered for each dataset.

Input FASTQ file (R2/second of pair)



12: R3G_REP3_R2.fastq
11: R3G_REP3_R1.fastq
10: R3G_REP2_R2.fastq
9: R3G_REP2_R1.fastq
8: R3G_REP1_R2.fastq

This is a batch mode input field. A separate job will be triggered for each dataset.

Perform initial ILLUMINACLIP step?

Yes

No

Agenda

- 9:00 Welcome
- 9:20 Basic Analysis with Galaxy
- 10:45 Break
- 11:00 Basic Analysis into Reusable Workflows
- 12:20 Lunch (on your own)
- 1:20 RNA-Seq Analysis, Part I
- 2:50 Break**
- 3:05 RNA-Seq Analysis, Part II
- 4:30 Launch your own Galaxy with AWS
- 5:00 Done

bit.ly/btigxy

Agenda

- 9:00 Welcome
- 9:20 Basic Analysis with Galaxy
- 10:45 Break
- 11:00 Basic Analysis into Reusable Workflows
- 12:20 Lunch (on your own)
- 1:20 RNA-Seq Analysis, Part I
- 2:50 Break
- 3:05 RNA-Seq Analysis, Part II**
- 4:30 Launch your own Galaxy with AWS
- 5:00 Done

bit.ly/btigxy

RNA-seq Exercise: Differential gene expression

Take samples under multiple conditions
(MeOH and R3G exposure in our example)

Map them

Count them

Compare them

RNA-Seq Mapping: Get the Data

Import into a new history:

Shared Data → Data Libraries → Training → RNA-Seq
→ UC-Davis* → Post QC reads → Still paired reads

Select first two

MeOH_REP1_R1 post QC

MeOH_REP1_R2 post QC

Shared Data → Data Libraries → Training → RNA-Seq

→ UC-Davis → Reference

Select GTF for hg38, chr12 from Sanger

* RNA-Seq example datasets from the 2013 UC Davis Bioinformatics Short Course. <http://bit.ly/ucdbsc2013>

RNA-seq Exercise: Mapping with Tophat2

- Tophat looks for best place(s) to map reads, and best places to insert introns
- *Imagine pages and pages of discussion on the intricacies and pitfalls of RNA-seq mapping here**

Mapping with Tophat: **mean inner distance**

Expected distance between paired end reads

- Determined by sample prep
- We'll use **90*** for **mean inner distance**
- We'll use **50** for **standard deviation**

* The library was constructed with the typical Illumina TruSeq protocol, which is supposed to have an average insert size of 200 bases. Our reads are 55 bases (R1) plus 55 bases (R2). So, the Inner Distance is estimated to be $200 - 55 - 55 = 90$

From the 2013 UC Davis Bioinformatics Short Course

Mapping with Tophat: **Use Existing Annotations?**

You can bias Tophat towards known annotations

- **Supply your own junction Data? → Yes**

- **Use Gene Annotation → Yes**

- **Gene Model Annotation →**

GTF for hg38, chr12 from Sanger

You can also restrict Tophat to known annotations

- **Use Raw Junctions → Yes** (tab delimited file)

- **Only look for supplied junctions → Yes**

Mapping with Tophat: **Make it quicker?**

Warning: Here be dragons!

- **Allow indel search** → **No**
- **Use Coverage Search** → **No** (wee dragons)

TopHat generates its database of possible splice junctions from two sources of evidence. The first and strongest source of evidence for a splice junction is when two segments from the same read (for reads of at least 45bp) are mapped at a certain distance on the same genomic sequence or when an internal segment fails to map - again suggesting that such reads are spanning multiple exons. With this approach, "GT-AG", "GC-AG" and "AT-AC" introns will be found *ab initio*. The second source is pairings of "coverage islands", which are distinct regions of piled up reads in the initial mapping. Neighboring islands are often spliced together in the transcriptome, so TopHat looks for ways to join these with an intron. **We only suggest users use this second option (--coverage-search) for short reads (< 45bp) and with a small number of reads (<= 10 million).** This latter option will only report alignments across "GT-AG" introns

Mapping w/ Tophat: **Max # of Alignments Allowed**

Some reads align to more than one place equally well.

For such reads, how many should Tophat include?

If more than the specified number, Tophat will pick those with the best mapping score.

Tophat **breaks ties randomly**.

Tophat assigns equal fractional credit to all n mappings

Instructs TopHat to allow up to this many alignments to the reference for a given read, and choose the alignments based on their alignment scores if there are more than this number. The default is 20 for read mapping. Unless you use `--report-secondary-alignments`, TopHat will report the alignments with the best alignment score. **If there are more alignments with the same score than this number, TopHat will randomly report only this many alignments.** In case of using `--report-secondary-alignments`, TopHat will try to report alignments up to this option value, and TopHat may randomly output some of the alignments with the same score to meet this number.

Mapping With Tophat: What to keep?


NGS BAM
Tools → Filter

This shows
two options
for cleanup.


Condition

1: Condition

Filter

1: Filter 


Select BAM property to filter on

mapQuality 


Filter on read mapping quality (phred scale)

>=20

You can use >, <, =, and ! (not) in your expression. E.g., to select reads with mapping quality of at least 30 use ">=30"

2: Filter 


Select BAM property to filter on


isProperPair 

Select properly paired reads

☒ Yes ☐ No

Checked = Read IS in proper pair, Empty = Read is NOT in the proper pair


 Insert Filter

 Insert Condition

Would you like to set rules?

☒ Yes ☐ No

Allows complex logical constructs. See Example 4 below.

 Execute

Only 5 more replicates to go!

Another way to avoid insanity is

Collections

RNA-Seq Differential Expression: Get the Data

Import into a new history:

Shared Data → Data Libraries → Training → RNA-Seq*

→ UC-Davis → Mapped

Select all (OK, maybe just half of them)

Shared Data → Data Libraries → Training → RNA-Seq*

→ UC-Davis → Reference

Select GTF for hg38, chr12 from Sanger




* RNA-Seq example datasets from the 2013 UC Davis Bioinformatics Short Course. <http://bit.ly/ucdbsc2013>


Dataset collections!




Dataset Collections give Galaxy **semantic knowledge about dataset relationships.**

Tools can then take advantage of this knowledge.




Dataset collections




History   




search datasets 




RNA-Seq thru Mapping, w
collections
12 shown
297.73 MB   




12: R3G REP3 R2 **Operations on multiple datasets**




11: R3G REP3 R1   




10: R3G REP2 R2   




9: R3G REP2 R1   




8: R3G REP1 R2   




7: R3G REP1 R1   




6: MeOH REP3 R2   




5: MeOH REP3 R1   


4: MeOH REP2 R2   




3: MeOH REP2 R1   

2: MeOH REP1 R2   

1: MeOH REP1 R1   

History   

search datasets 

RNA-Seq thru Mapping, w
collections
12 shown
297.73 MB   

All None For all selected...

☐ Hide datasets
☐ Unhide datasets
☐ Delete datasets
☐ Undelete datasets
☐ Permanently delete datasets
☐ Build Dataset List
☐ Build Dataset Pair
☒ Build List of Dataset Pairs

☒ 6: MeOH REP3 R2

☒ 5: MeOH REP3 R1

☒ 4: MeOH REP2 R2

☒ 3: MeOH REP2 R1

☒ 2: MeOH REP1 R2

☒ 1: MeOH REP1 R1

Create a collection of paired datasets

Could not automatically create any pairs from the given dataset names

0 unpaired forward – (6 filtered out)

_1

[Choose filters](#) [Clear filters](#)

[Auto-pair](#)

Choose from the following filters to change which unpaired reads are shown in the display:

Forward: _1, Reverse: _2

Forward: _R1, Reverse: _R2

0 unpaired reverse – (6 filtered out)

_2

Analyze Data

Workflow

Shared Data

Visualization

Admin

Help

User

Create a collection of paired datasets

Could not automatically create any pairs from the given dataset names

3 unpaired forward – (3 filtered out)

_R1

[Choose filters](#) [Clear filters](#)

[Auto-pair](#)

Pair these datasets

Pair these datasets

Pair these datasets

3 unpaired reverse – (3 filtered out)

_R2

MeOH_REP1_R2

MeOH_REP2_R2

MeOH_REP3_R2

MeOH_REP1_R1

MeOH_REP2_R1

MeOH_REP3_R1

Create a collection of paired datasets

3 pairs created: all datasets have been successfully paired

0 unpaired forward – (0 filtered out) Choose filters Clear filters 0 unpaired reverse – (0 filtered out)

_R1 _R2

3 paired Unpair all




| | | | |
|----------------|-----------|----------------|---|
| MeOH_REP1_R1 → | MeOH_REP1 | ← MeOH_REP1_R2 | 🔗 |
| MeOH_REP2_R1 → | MeOH_REP2 | ← MeOH_REP2_R2 | 🔗 |
| MeOH_REP3_R1 → | MeOH_REP3 | ← MeOH_REP3_R2 | 🔗 |


Remove file extensions from pair names? ☒




Name: MeOH

Cancel Create list

Dataset collections

History   

search datasets 

RNA-Seq thru Mapping, w
collections
14 shown
297.73 MB   

All None **Operations on multiple
datasets**

☐ **14: R3G**
a list of paired datasets

☐ **13: MeOH**
a list of paired datasets

☒ **12: R3G REP3 R2**

☒ **11: R3G REP3 R1**

☒ **10: R3G REP2 R2**

☒ **9: R3G REP2 R1**

☒ **8: R3G REP1 R2**

☒ **7: R3G REP1 R1**

☐ **6: MeOH REP3 R2**




☐ **5: MeOH REP3 R1**

☐ **4: MeOH REP2 R2**

☐ **3: MeOH REP2 R1**

☐ **2: MeOH REP1 R2**

☐ **1: MeOH REP1 R1**

History   




[Back to RNA-Seq thru Mapping, w
collections](#)

MeOH
a list of paired datasets

MeOH REP1
a pair of datasets



MeOH REP2
a pair of datasets



MeOH REP3
a pair of datasets

History   

[Back to MeOH](#)

MeOH_REP1
a pair of datasets

forward  

reverse  

Dataset collections Created

History

search datasets

RNA-Seq thru Mapping, w
collections
14 shown
297.73 MB

☒

All

None

Operations on multiple
datasets

☐

14: R3G
a list of paired datasets

☐

13: MeOH
a list of paired datasets

☒

12: R3G REP3 R2

☒

11: R3G REP3 R1

☒

10: R3G REP2 R2

☒

9: R3G REP2 R1

☒

8: R3G REP1 R2

☒

7: R3G REP1 R1

☐

6: MeOH REP3 R2

☐

5: MeOH REP3 R1

☐

4: MeOH REP2 R2

☐

3: MeOH REP2 R1

☐

2: MeOH REP1 R2

☐

1: MeOH REP1 R1

Differential expression with CuffDiff

Part of the Tuxedo RNA-Seq Suite (as are Tophat, Bowtie, StringTie, Cufflinks, Cuffmerge, ...)

Identifies differential expression between multiple datasets

Widely used and widely installed on Galaxy instances

NGS: RNA Analysis → Cuffdiff

Cuffdiff

Cuffdiff previously used FPKM/RPKM as central statistic.

Total # mapped reads heavily influences FPKM/RPKM.
Can lead to challenges when you have very highly expressed genes in the mix.

Now supports geometric normalization, the same model used by DESeq (and in fact, it's now the default). Less prone to distortion from highly expressed genes.

Cuffdiff: Which transcript definitions to use?

We'll use the official genome annotations

But there are a world of options out there for discovering and using novel transcripts.

StringTie, Cufflinks, Cuffmerge, ...

Cuffdiff

- Running with 2 Groups: MeOH and R3G
- Each group has 3 replicates each
- Can take advantage of collections

Transcripts

13: GTF for hg38, chr12 from Sanger



A transcript GFF3 or GTF file produced by cufflinks, cuffcompare, or other source.

Omit Tabular Datasets

Yes

No

Discard the tabular output.

Generate SQLite

Yes

No

Generate a SQLite database for use with cummeRbund.

Input data type

SAM/BAM



CuffNorm supports either CXB (from cuffquant) or SAM/BAM input files. Mixing is not supported. Default: SAM/BAM

Condition**1: Condition****Name**

MeOH

Replicates

62: HISAT2 on R3G

**2: Condition****Name**

R3G

Replicates

58: HISAT2 on MeOH



Cuffdiff

Execute it

Cuffdiff

Produces many output files, all explained in doc

We'll focus on **gene differential expression testing**

| test_id | gene_id | gene | locus | sample_1 | sample_2 | status | value_1 | value_2 | log2(fold_change) | test_stat | p_value | q_value | significant |
|----------|----------|----------|---------------------------|----------|----------|--------|----------|----------|-------------------|-----------|---------|-------------|-------------|
| A2M | A2M | A2M | chr12:9217772-9268558 | MeOH | R3G | NOTEST | 3.32147 | 3.13694 | -0.0824644 | 0 | 1 | 1 | no |
| A2M-AS1 | A2M-AS1 | A2M-AS1 | chr12:9217772-9268558 | MeOH | R3G | NOTEST | 7.45797 | 13.9413 | 0.902515 | 0 | 1 | 1 | no |
| A2ML1 | A2ML1 | A2ML1 | chr12:8975149-9029381 | MeOH | R3G | NOTEST | 4.83055 | 7.79884 | 0.691072 | 0 | 1 | 1 | no |
| A2MP1 | A2MP1 | A2MP1 | chr12:9381128-9386803 | MeOH | R3G | NOTEST | 2.49656 | 0 | -inf | 0 | 1 | 1 | no |
| AAAS | AAAS | AAAS | chr12:53701239-53715412 | MeOH | R3G | OK | 269.035 | 159.23 | -0.756683 | -2.22857 | 0.0005 | 0.00194017 | yes |
| AACS | AACS | AACS | chr12:125549924-125627871 | MeOH | R3G | NOTEST | 29.2933 | 35.0339 | 0.258178 | 0 | 1 | 1 | no |
| ABCB9 | ABCB9 | ABCB9 | chr12:123405497-123451056 | MeOH | R3G | NOTEST | 4.68869 | 1.7732 | -1.40283 | 0 | 1 | 1 | no |
| ABCC9 | ABCC9 | ABCC9 | chr12:21950323-22089628 | MeOH | R3G | OK | 553.247 | 487.261 | -0.18323 | -2.02806 | 0.0004 | 0.00162143 | yes |
| ABCD2 | ABCD2 | ABCD2 | chr12:39945021-40013843 | MeOH | R3G | OK | 86.1377 | 172.795 | 1.00435 | 4.3436 | 5e-05 | 0.000246739 | yes |
| ACACB | ACACB | ACACB | chr12:109577201-109706030 | MeOH | R3G | NOTEST | 8.45306 | 15.5772 | 0.881885 | 0 | 1 | 1 | no |
| ACAD10 | ACAD10 | ACAD10 | chr12:112123856-112194911 | MeOH | R3G | NOTEST | 21.8237 | 27.8326 | 0.350882 | 0 | 1 | 1 | no |
| ACADS | ACADS | ACADS | chr12:121163570-121177811 | MeOH | R3G | NOTEST | 38.644 | 16.1739 | -1.25658 | 0 | 1 | 1 | no |
| ACRBP | ACRBP | ACRBP | chr12:6747241-6756580 | MeOH | R3G | NOTEST | 2.96987 | 3.26939 | 0.138621 | 0 | 1 | 1 | no |
| ACSM4 | ACSM4 | ACSM4 | chr12:7456927-7480969 | MeOH | R3G | NOTEST | 0 | 0 | 0 | 0 | 1 | 1 | no |
| ACSS3 | ACSS3 | ACSS3 | chr12:81471808-81649582 | MeOH | R3G | NOTEST | 0 | 0 | 0 | 0 | 1 | 1 | no |
| ACTR6 | ACTR6 | ACTR6 | chr12:100593864-100618202 | MeOH | R3G | OK | 475.594 | 421.324 | -0.174799 | -0.797581 | 0.1588 | 0.258406 | no |
| ACVR1B | ACVR1B | ACVR1B | chr12:52345450-52390863 | MeOH | R3G | NOTEST | 32.5737 | 38.3075 | 0.233922 | 0 | 1 | 1 | no |
| ACVRL1 | ACVRL1 | ACVRL1 | chr12:52301201-52317145 | MeOH | R3G | NOTEST | 1.27713 | 2.16161 | 0.759201 | 0 | 1 | 1 | no |
| ADAM1A | ADAM1A | ADAM1A | chr12:112336866-112339706 | MeOH | R3G | NOTEST | 30.0162 | 55.2154 | 0.879331 | 0 | 1 | 1 | no |
| ADAMTS20 | ADAMTS20 | ADAMTS20 | chr12:43748011-43945724 | MeOH | R3G | NOTEST | 0.453322 | 0.502067 | 0.147346 | 0 | 1 | 1 | no |
| ADCY6 | ADCY6 | ADCY6 | chr12:49159974-49182820 | MeOH | R3G | NOTEST | 9.32722 | 17.6743 | 0.922135 | 0 | 1 | 1 | no |
| ADIPOR2 | ADIPOR2 | ADIPOR2 | chr12:1800246-1897845 | MeOH | R3G | OK | 207.468 | 179.333 | -0.210248 | -1.02392 | 0.09 | 0.158988 | no |
| AEBP2 | AEBP2 | AEBP2 | chr12:19592607-19675173 | MeOH | R3G | OK | 143.039 | 128.293 | -0.156957 | -0.688267 | 0.2254 | 0.344537 | no |
| AGAP2 | AGAP2 | AGAP2 | chr12:58118075-58135944 | MeOH | R3G | OK | 98.2385 | 116.302 | 0.243511 | 0.935119 | 0.11475 | 0.198086 | no |
| AICDA | AICDA | AICDA | chr12:8754761-8765442 | MeOH | R3G | NOTEST | 78.1514 | 63.4313 | -0.301077 | 0 | 1 | 1 | no |
| AKAP3 | AKAP3 | AKAP3 | chr12:4724675-4754343 | MeOH | R3G | NOTEST | 6.12385 | 7.89626 | 0.366731 | 0 | 1 | 1 | no |
| ALDH1L2 | ALDH1L2 | ALDH1L2 | chr12:105413561-105478341 | MeOH | R3G | NOTEST | 7.11374 | 8.11722 | 0.190377 | 0 | 1 | 1 | no |
| ALDH2 | ALDH2 | ALDH2 | chr12:112204690-112247789 | MeOH | R3G | NOTEST | 12.8033 | 8.05635 | -0.668321 | 0 | 1 | 1 | no |
| ALG10 | ALG10 | ALG10 | chr12:34175215-34181236 | MeOH | R3G | NOTEST | 54.8575 | 59.3459 | 0.11346 | 0 | 1 | 1 | no |
| ALG10B | ALG10B | ALG10B | chr12:38710556-38723528 | MeOH | R3G | NOTEST | 43.8157 | 63.0457 | 0.524952 | 0 | 1 | 1 | no |
| ALKBH2 | ALKBH2 | ALKBH2 | chr12:109525992-109531293 | MeOH | R3G | OK | 679.517 | 297.183 | -1.19316 | -3.34255 | 5e-05 | 0.000246739 | yes |
| ALX1 | ALX1 | ALX1 | chr12:85674035-85695561 | MeOH | R3G | NOTEST | 0 | 0 | 0 | 0 | 1 | 1 | no |

Cuffdiff: differentially expressed genes

| Column | Contents |
|-------------|---|
| test_stat | value of the test statistic used to compute significance of the observed change |
| p_value | Uncorrected P value for test statistic |
| q_value | FDR-adjusted p-value for the test statistic |
| status | Was there enough data to run the test? |
| significant | and, was the gene differentially expressed? |

Cuffdiff

- Column 7 ("status") can be FAIL, NOTEST, LOWDATA or OK
 - Filter and Sort → Filter
 - `c7 == 'OK'`
- Column 14 ("significant") can be yes or no
 - Filter and Sort → Filter
 - `c14 == 'yes'`

Returns the list of genes with

- 1) enough data to make a call, and
- 2) that are called as differentially expressed.

Cuffdiff: Next Steps

Try running Cuffdiff with different **normalization** and **dispersion estimation** methods.

Compare the differentially expressed gene lists.
Which settings have what type of impacts on the results?

Are there any patterns to the identified genes?

Agenda

- 9:00 Welcome
- 9:20 Basic Analysis with Galaxy
- 10:45 Break
- 11:00 Basic Analysis into Reusable Workflows
- 12:20 Lunch (on your own)
- 1:20 RNA-Seq Analysis, Part I
- 2:50 Break
- 3:05 RNA-Seq Analysis, Part II
- 4:30 Launch your own Galaxy with AWS
- 5:00 Done

bit.ly/btigxy

<https://launch.usegalaxy.org/>

Agenda

- 9:00 Welcome
- 9:20 Basic Analysis with Galaxy
- 10:45 Break
- 11:00 Basic Analysis into Reusable Workflows
- 12:20 Lunch (on your own)
- 1:20 RNA-Seq Analysis, Part I
- 2:50 Break
- 3:05 RNA-Seq Analysis, Part II
- 4:30 Launch your own Galaxy with AWS
- 5:00 Done**

bit.ly/btigxy

Acknowledgements

| | |
|---------------|--------------------------|
| You | AWS |
| Surya Saha | |
| Lukas Mueller | NIH |
| John Ashton | Johns Hopkins University |
| | Penn State University |

Boyce Thompson Institute
Cornell University

bit.ly/btigxy_feedback



Thanks

Agenda

- 9:00 Welcome
- 9:20 Basic Analysis with Galaxy
- 10:45 Break
- 11:00 Basic Analysis into Reusable Workflows
- 12:20 Lunch (on your own)
- 1:20 RNA-Seq Analysis, Part I
- 2:50 Break
- 3:05 RNA-Seq Analysis, Part II
- 4:30 Launch your own Galaxy with AWS
- 5:00 Done

bit.ly/btigxy_feedback

bit.ly/btigxy_feedback

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



Le Corum
Conference centre

gcc2017.sciencesconf.org

November 7-11



Salt Lake City, Utah

Galaxy Community Resources: Galaxy **Biostar**

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

Absolutely have to **encourage community support**.

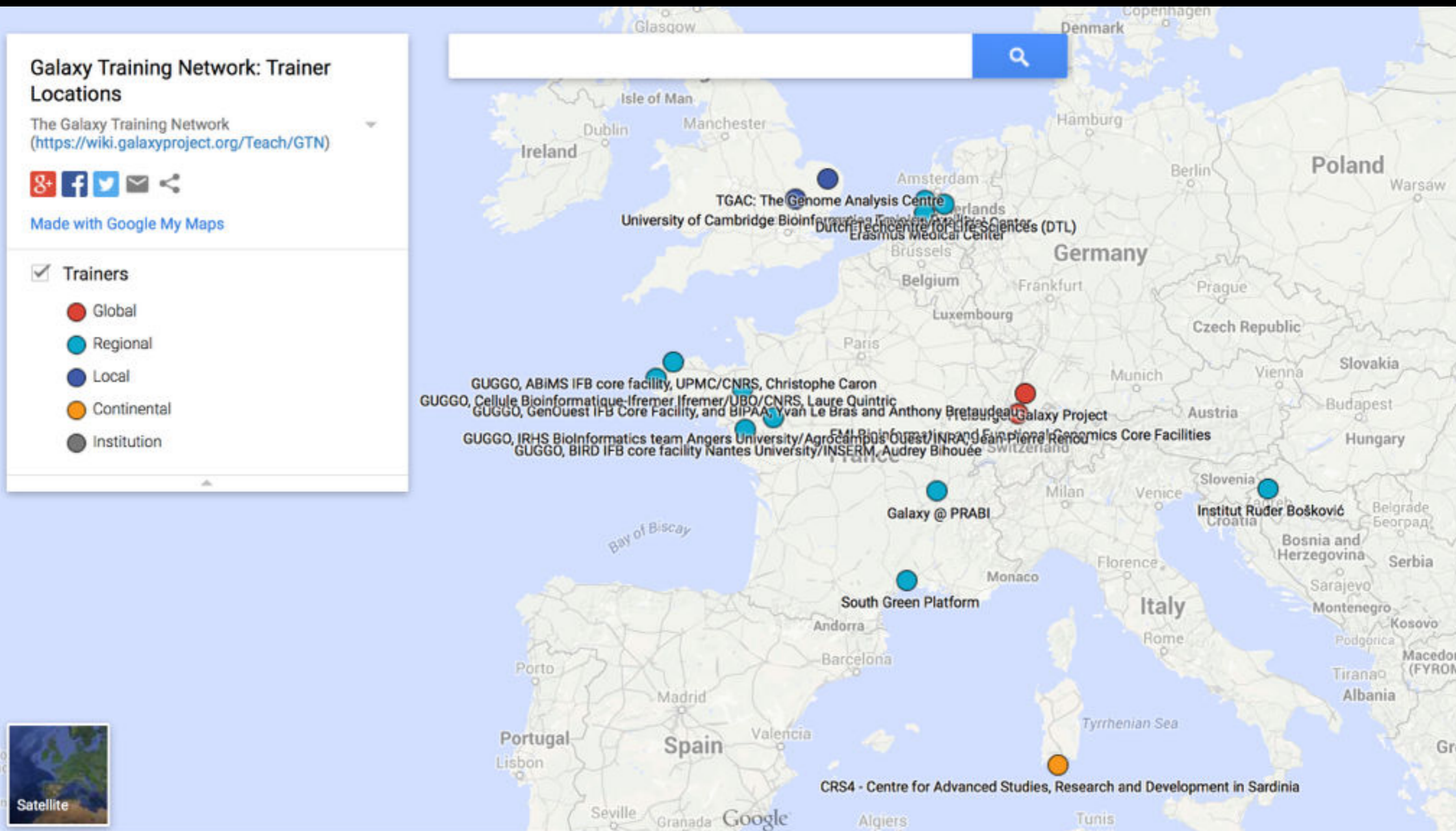
Project traditionally used mailing list

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

Galaxy Community Resources: Mailing Lists

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

Galaxy-Dev

Questions about developing for and deploying Galaxy

High volume (2336 posts in 2015, 1000+ members)

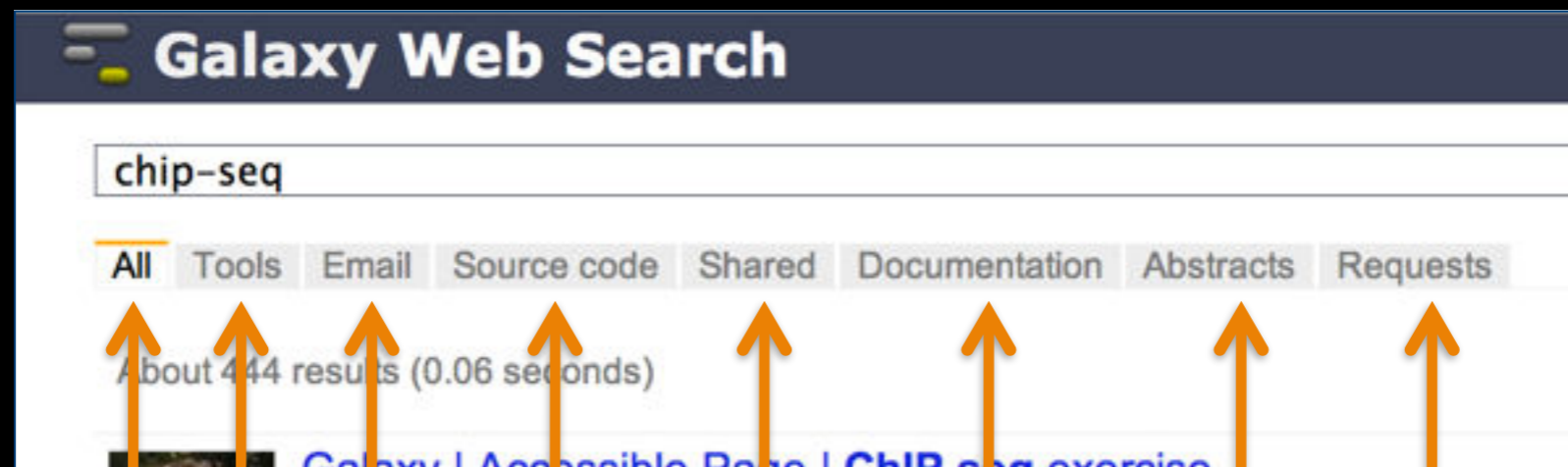
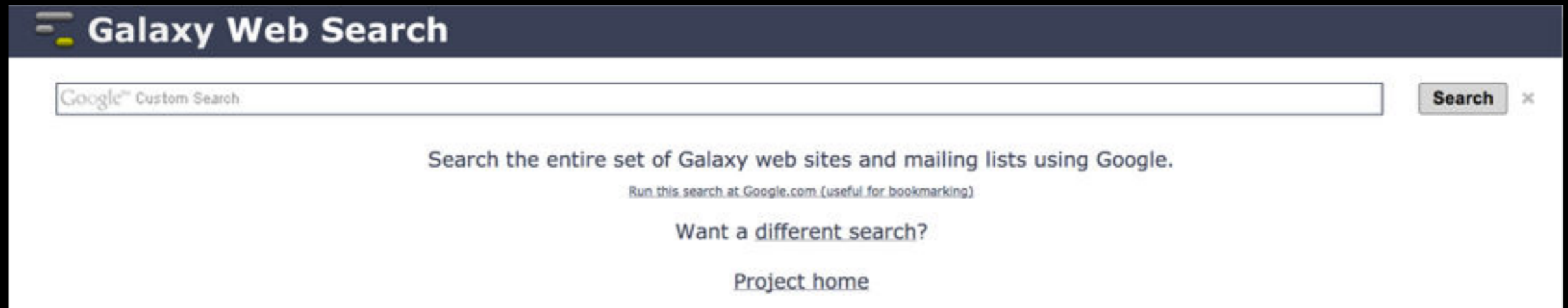
Galaxy-Announce

Project announcements, low volume, moderated

Low volume (36 posts in 2015, 6500+ members)

Also Galaxy-UK, -France, -Proteomics, -Training, ...

Unified Search: <http://galaxyproject.org/search>



Find

Everything on ...

Tools for ...

Email about ...

Source code for ...

Published Histories, Pages, Workflows, about ...

Documentation on ...

Papers using Galaxy for ...

Related feature requests



Galaxy is an open, web-based platform for *accessible, reproducible, and transparent* computational biomedical research.

- **Accessible:** Users without programming experience can easily specify parameters and run tools and workflows.
- **Reproducible:** Galaxy captures information so that any user can repeat and understand a complete computational analysis.
- **Transparent:** Users share and publish analyses via the web and create Pages, interactive, web-based documents that describe a complete analysis.

This is the Galaxy Community Wiki. It describes all things Galaxy.

Use Galaxy

Galaxy's public web server usegalaxy.org makes analysis tools, genomic data, tutorial demonstrations, persistent workspaces, and publication services available to any scientist. Extensive [user documentation](#) applicable to any [public](#) or local Galaxy instance is available.



Community & Project

Galaxy has a large and active user community and many ways to get involved.

- [Community](#)

Deploy Galaxy

Galaxy is a free and open source project available to all. Local Galaxy servers can be set up by [downloading](#) the Galaxy application.

- [Admin](#)
- [Cloud](#)



Contribute

- **Users:** [Share](#) your histories, workflows, visualizations, data libraries, and [Galaxy Pages](#), enabling others to use and learn from them.



Use Galaxy

[Servers](#) • [Learn Main](#) • [Choices](#)
[Share](#) • [Search](#)

Communicate

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

Deploy Galaxy

[Get Galaxy](#) • [Cloud Admin](#) • [Tool Config](#)
[Tool Shed](#) • [Search](#)

Contribute

[Develop](#) • [Tools](#)
[Issues & Requests](#)
[Logs](#) • [Deployments](#)
[Teach](#)

Galaxy Project

[Home](#) • [About](#) • [Cite Community](#)
[Big Picture](#)

Events

News

[DaveClements](#)
[Settings](#)
[Logout](#)
 |
 Search:

[Titles](#)
[Text](#)

[Events](#)
[Edit](#)
[History](#)
[Actions](#)

Galaxy Event Horizon

Events with Galaxy-related content are listed here.

Also see the [Galaxy Events Google Calendar](#) for a listing of events and deadlines that are in the Galaxy Community. This is also available as an [RSS feed](#).

If you know of any event that should be added to this page and/or to the Galaxy Event Calendar, send it to outreach@galaxyproject.org.

For events prior to this year, see the [Events Archive](#).

Upcoming Events

| Date | Topic/Event | Venue/Location |
|----------------|---|---|
| December 12 | Introduction to Galaxy Workshop | Virginia State University, Petersburg, Virginia |
| December 16-19 | RNA-Seq and ChIP-Seq Analysis with Galaxy | UC Davis, California, United States |
| 2015 | | |
| January 10-14 | Galaxy for SNP and Variant Data Analysis | Plant and Animal Genome XXIII (PAG2014), States |
| January 19-20 | NGS pipelines with Galaxy | e-Infrastructures for Massively Parallel Sequencing, Sweden |
| February 9-13 | Analyse bioinformatique de séquences sous Galaxy | Montpellier, France |
| February 16-18 | Accessible and Reproducible Large-Scale Analysis with Galaxy | Genome and Transcriptome Analysis, Pacific Conference, San Francisco, California |
| | Large-Scale NGS data Analysis on Amazon Web Services Using Globus Genomic | Genomics & Sequencing Data Integration, of Molecular Medicine Tri-Conference, San Francisco, California |

News Items

Opening at McMaster University

The [McArthur Lab](#) in the [McMaster University Department of Biochemistry & Biomedical Sciences](#) is seeking a Systems Administrator / Information Technologist to help establish a new bioinformatics laboratory at McMaster, plus develop the next generation of the [Comprehensive Antibiotic Resistance Database \(CARD\)](#).

From the [job announcement on Evoldir](#):

The candidate will configure BLADE and other hardware for general bioinformatics analysis, development of a GIT version control system, **construction of an in house Galaxy server (usegalaxy.org)**, and development of a new interface, stand-alone tools, APIs, and algorithms for the CARD (based on [Chado](#)).

See the [full announcement](#) for details.

Posted to the [Galaxy News](#) on 2014-12-05

December 2014 Galaxy Newsletter

As always there's a lot going on in the Galaxy this month. "Like what?" you say. Well, read the dang [December Galaxy Newsletter](#) we say! Highlights include:

- [Galaxy Day! In Paris! This Wednesday!](#)
- Near Richmond, Virginia? There's a [Galaxy Workshop at Virginia State U on December 12](#).
- [GCC2015 needs sponsors!](#)
- Other [upcoming events](#) on two continents
- **96 new papers**, including 6 highlighted papers, referencing, using, extending, and implementing Galaxy.
- [Job openings at 7+ organizations](#)
- A new mailing list: [Galaxy-Training](#)
- [15 new ToolShed repositories](#) from 10 contributors
- And, 10 other juicy (well maybe not *juicy*, but certainly not *crunchy*) [bits of news](#)

Dave Clements and the *crisp* Galaxy Team

Posted to the [Galaxy News](#) on 2014-12-01

Bioinformaticians, Freiburg

[Max Planck Institute of Immunobiology and Epigenetics](#) in Freiburg, Germany has an opening for a Bioinformatician for an initial period of two years. The successful candidate will work at the interface between an in-house deep-sequencing facility (HiSeq-2500) and the various research groups at the institute. Main responsibilities include

primary analysis of deep-sequencing data and quality control

Galaxy Resources & Community: Videos

The screenshot shows the Vimeo profile for the 'Galaxy Project'. The header includes the Vimeo logo and navigation links: Me, Videos, Create, Watch, Tools, Upload. A search bar is on the right. The profile name 'Galaxy Project' is followed by a 'PLUS' badge and the text 'Joined 1 month ago'. Below this is a video player showing a dark interface with three buttons. To the right of the player are statistics: 54 Videos, 0 Likes, 0 Following, 1 Group, 6 Channels, and 0 Albums. A 'Recently Uploaded' section follows, with a link to 'See all 54 videos'. It displays four video thumbnails: 'Using Galaxy protocol 3: Calling Peaks For ChIP-seq Data' (CPB Using Galaxy 3, 5 days ago), 'Using Galaxy protocol 2: Loading Data and Understanding Datatypes' (CPB Using Galaxy 2, 5 days ago), 'Using Galaxy protocol 1: Finding Human Coding Exons with Highest SNP Density' (CPB Using Galaxy 1, 5 days ago), and 'FASTQ Prep Illumina' (FASTQ Prep - Illumina, 1 week ago). A 'Settings' button is located below the video player. A descriptive paragraph about the Galaxy project is at the bottom left of the page.

Galaxy Project PLUS
Joined 1 month ago

54 Videos, 0 Likes, 0 Following, 1 Group, 6 Channels, 0 Albums

Recently Uploaded + See all 54 videos

- Using Galaxy protocol 3: Calling Peaks For ChIP-seq Data (CPB Using Galaxy 3, 5 days ago)
- Using Galaxy protocol 2: Loading Data and Understanding Datatypes (CPB Using Galaxy 2, 5 days ago)
- Using Galaxy protocol 1: Finding Human Coding Exons with Highest SNP Density (CPB Using Galaxy 1, 5 days ago)
- FASTQ Prep Illumina (FASTQ Prep - Illumina, 1 week ago)

Settings

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for

“How to”
screencasts on
using and
deploying
Galaxy

Talks from
previous
meetings.

<http://vimeo.com/galaxyproject>

Galaxy Resources & Community: CiteULike Group

Now
almost
3000
papers



CiteULike Group: Galaxy Search Register Log in

Group: Galaxy - library 2336 articles

Search Copy Export Sort Hide Details

✓ Adaptation of the targeted capture Methyl-Seq platform for the mouse genome identifies novel tissue-specific methylation patterns of genes involved in neurodevelopment

Epigenetics (18 May 2015), pp. 00-00, doi:10.1080/15592294.2015.1045179
by Benjamin Hing, Enrique Ramos, Patricia Braun, et al.
posted to methods by galaxyproject to the group Galaxy on 2015-05-28 21:46:38 ★★

■ Abstract

✓ Genomic and experimental evidence for multiple metabolic functions in the RidA/YjgF/YER057c/U

BMC Genomics, Vol. 16, No. 1. (15 May 2015), 382, doi:10.1186/s12864-015-1584-3
by Thomas D. Niehaus, Svetlana Gerdes, Kelsey Hodge-Hanson, et al.
posted to methods usemain by galaxyproject to the group Galaxy on 2015-05-28 21:41:14 ★★

■ Abstract

✓ NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data

Nat. Protocols, Vol. 10, No. 6. (07 June 2015), pp. 823-844, doi:10.1038/nprot.2015.052
by Jianguo Xia, Erin E. Gill, Robert E. W. Hancock
posted to visualization by galaxyproject to the group Galaxy on 2015-05-28 21:37:43 ★★ along with 2 people and

✓ Repression by H-NS of genes required for the biosynthesis of the Vibrio cholerae biofilm matrix is mediated by the

Molecular Microbiology (1 May 2015), pp. n/a-n/a, doi:10.1111/mmi.13058
by Julio C. Ayala, Hongxia Wang, Anisia J. Silva, Jorge A. Benitez
posted to methods usemain by galaxyproject to the group Galaxy on 2015-05-28 21:30:30 ★★

■ Abstract

✓ A Sleeping Beauty forward genetic screen identifies new genes and pathways driving osteosarcoma development and

| Group Tags | |
|------------------------------------|------|
| All tags in the group Galaxy | |
| Filter: | |
| [Display as Cloud] | |
| methods | 1149 |
| workbench | 702 |
| usemain | 233 |
| tools | 169 |
| usepublic | 129 |
| isgalaxy | 124 |
| uselocal | 90 |
| cloud | 89 |
| shared | 81 |
| other | 68 |
| refpublic | 57 |
| unknown | 53 |
| reproducibility | 51 |
| howto | 45 |
| project | 43 |
| visualization | 15 |
| usecloud | 4 |

<http://bit.ly/gxycul>

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>

A free for everyone web service:

<http://usegalaxy.org>

A free (for everyone) web server integrating a wealth of tools, compute resources, petabytes of reference data and permanent storage



CYVERSE™

However, *a centralized solution cannot support the different analysis needs of the entire world.*

What is Galaxy?

**Data integration and analysis platform that
emphasizes accessibility, reproducibility, and
transparency**

<http://galaxyproject.org>

Yay! We have a list of genes and overlap counts!*

Now, what can we do with that?

All sorts of things.

* Technically, we have a list of gene symbols, and the maximum number of overlapping repeats from any of its transcripts. We also haven't done things like normalize the scores based on gene length. Your mileage may vary. Let's not sweat the details.

GO Term Enrichment

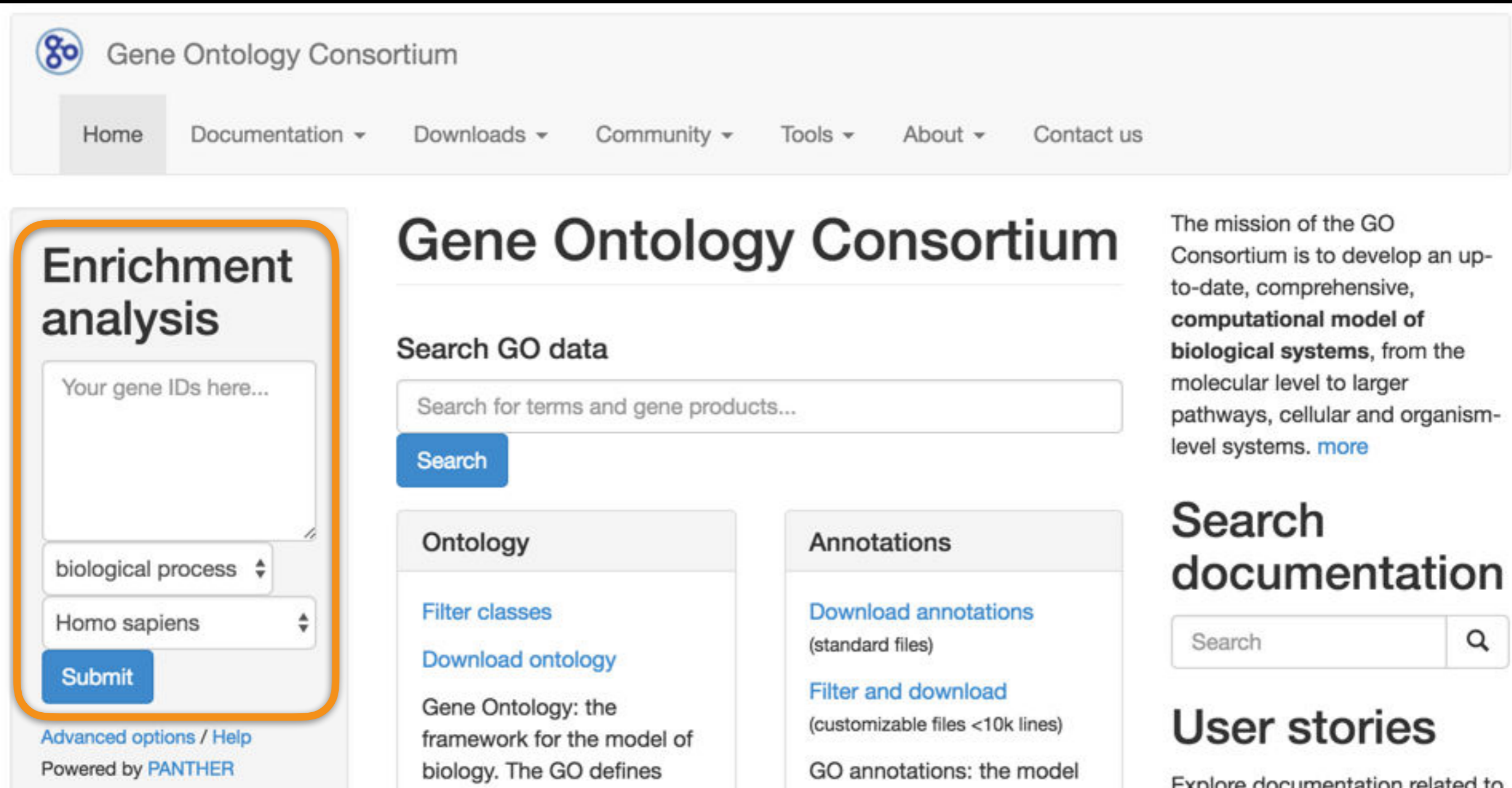
Do genes with particular functions tend to occur in this list more often than they would by random chance?



GO: Create a list of just the gene symbols

Remember how?

(Stop or) GO: Can do this step, or just watch



The screenshot shows the Gene Ontology Consortium website. The top navigation bar includes links for Home, Documentation, Downloads, Community, Tools, About, and Contact us. The main header reads "Gene Ontology Consortium". On the left, a sidebar titled "Enrichment analysis" is highlighted with an orange border. It contains a text input field for "Your gene IDs here...", two dropdown menus for "biological process" and "Homo sapiens", and a "Submit" button. Below the sidebar, there are links for "Advanced options / Help" and "Powered by PANTHER". The main content area features a "Search GO data" section with a search bar and a "Search" button. Below this are two columns: "Ontology" with links for "Filter classes" and "Download ontology", and "Annotations" with links for "Download annotations (standard files)" and "Filter and download (customizable files <10k lines)". To the right of the main content, there is a mission statement, a "Search documentation" section with a search bar, and a "User stories" section.

Gene Ontology Consortium

Home Documentation Downloads Community Tools About Contact us

Enrichment analysis

Your gene IDs here...

biological process

Homo sapiens

Submit

[Advanced options / Help](#)
Powered by [PANTHER](#)

Gene Ontology Consortium

Search GO data

Search for terms and gene products...

Search

Ontology

[Filter classes](#)

[Download ontology](#)

Gene Ontology: the framework for the model of biology. The GO defines

Annotations

[Download annotations](#)
(standard files)

[Filter and download](#)
(customizable files <10k lines)

GO annotations: the model

The mission of the GO Consortium is to develop an up-to-date, comprehensive, **computational model of biological systems**, from the molecular level to larger pathways, cellular and organism-level systems. [more](#)

Search documentation

Search

User stories

Explore documentation related to

<http://geneontology.org/>

GO: Results from whole genome, 1 or more overlapping repeats (8969 genes)

Displaying only results with P<0.05; [click here to display all results](#)

| | Homo sapiens (REF) | upload 1 (▼ Hierarchy NEW! ?) | | | | |
|---|-----------------------|-------------------------------|----------|-----------------|-----|----------|
| GO biological process complete | # | # | expected | Fold Enrichment | +/- | P value |
| chromatin modification | 289 | 196 | 123.50 | 1.59 | + | 6.66E-06 |
| ↳ chromatin organization | 636 | 376 | 271.78 | 1.38 | + | 5.60E-06 |
| ↳ chromosome organization | 984 | 555 | 420.49 | 1.32 | + | 6.19E-07 |
| ↳ organelle organization | 3133 | 1636 | 1338.83 | 1.22 | + | 4.94E-14 |
| ↳ cellular component organization | 5133 | 2606 | 2193.49 | 1.19 | + | 1.25E-19 |
| ↳ cellular process | 14559 | 6671 | 6221.52 | 1.07 | + | 4.95E-22 |
| ↳ cellular component organization or biogenesis | 5288 | 2688 | 2259.73 | 1.19 | + | 7.21E-21 |
| ↳ macromolecular complex subunit organization | 1983 | 1021 | 847.40 | 1.20 | + | 4.89E-06 |
| peptidyl-lysine modification | 314 | 194 | 134.18 | 1.45 | + | 4.87E-03 |
| ↳ peptidyl-amino acid modification | 855 | 456 | 365.37 | 1.25 | + | 1.34E-02 |
| ↳ cellular protein modification process | 2836 | 1397 | 1211.91 | 1.15 | + | 9.10E-05 |
| ↳ protein modification process | 2836 | 1397 | 1211.91 | 1.15 | + | 9.10E-05 |
| ↳ protein metabolic process | 4036 | 1908 | 1724.71 | 1.11 | + | 5.29E-03 |
| ↳ macromolecule metabolic process | 7359 | 3685 | 3144.73 | 1.17 | + | 1.39E-28 |
| ↳ organic substance metabolic process | 9032 | 4308 | 3859.66 | 1.12 | + | 7.16E-18 |
| ↳ metabolic process | 9443 | 4480 | 4035.29 | 1.11 | + | 2.03E-17 |
| ↳ primary metabolic process | 8601 | 4133 | 3675.48 | 1.12 | + | 6.53E-19 |
| ↳ macromolecule modification | 3007 | 1480 | 1284.99 | 1.15 | + | 3.58E-05 |

Published History: Gene-Repeat overlap, entire genome

GO: Results from whole genome, 2 or more overlapping repeats (2759 genes)

Displaying only results with P<0.05; [click here to display all results](#)

| | Homo sapiens (REF) | upload 1 (▼ Hierarchy NEW! ?) | | | | |
|---|-----------------------|--|----------|-----------------|-----|----------|
| GO biological process complete | # | # | expected | Fold Enrichment | +/- | P value |
| membrane depolarization during action potential | 39 | 19 | 5.22 | 3.64 | + | 2.01E-02 |
| ↳ biological regulation | 11384 | 1776 | 1523.15 | 1.17 | + | 2.20E-18 |
| ↳ membrane depolarization | 61 | 24 | 8.16 | 2.94 | + | 3.99E-02 |
| regulation of histone methylation | 59 | 25 | 7.89 | 3.17 | + | 7.12E-03 |
| ↳ regulation of histone modification | 129 | 43 | 17.26 | 2.49 | + | 9.63E-04 |
| ↳ regulation of primary metabolic process | 5720 | 1046 | 765.32 | 1.37 | + | 5.02E-27 |
| ↳ regulation of metabolic process | 6087 | 1096 | 814.43 | 1.35 | + | 2.49E-26 |
| ↳ regulation of biological process | 10767 | 1708 | 1440.60 | 1.19 | + | 1.66E-20 |
| ↳ regulation of macromolecule metabolic process | 5730 | 1052 | 766.66 | 1.37 | + | 6.06E-28 |
| ↳ regulation of cellular metabolic process | 5781 | 1058 | 773.48 | 1.37 | + | 1.21E-27 |
| ↳ regulation of cellular process | 10292 | 1651 | 1377.04 | 1.20 | + | 1.75E-21 |
| ↳ regulation of chromatin organization | 152 | 50 | 20.34 | 2.46 | + | 1.43E-04 |
| ↳ regulation of chromosome organization | 272 | 66 | 36.39 | 1.81 | + | 4.52E-02 |
| ↳ regulation of organelle organization | 1097 | 211 | 146.78 | 1.44 | + | 1.37E-03 |
| ↳ regulation of cellular component organization | 2246 | 409 | 300.51 | 1.36 | + | 1.22E-06 |
| histone lysine methylation | 64 | 27 | 8.56 | 3.15 | + | 2.90E-03 |
| ↳ histone methylation | 84 | 31 | 11.24 | 2.76 | + | 6.88E-03 |
| ↳ histone modification | 337 | 88 | 45.09 | 1.95 | + | 5.91E-05 |
| ↳ covalent chromatin modification | 346 | 92 | 46.29 | 1.99 | + | 1.16E-05 |
| ↳ macromolecule metabolic process | 7359 | 1260 | 984.62 | 1.28 | + | 4.66E-23 |
| ↳ organic substance metabolic process | 9032 | 1384 | 1208.46 | 1.15 | + | 1.23E-07 |

GO: Results from whole genome, 3 or more overlapping repeats (986 genes)

Displaying only results with P<0.05; [click here to display all results](#)

| | Homo sapiens (REF) | upload_1 (▼ Hierarchy NEW! ?) | | | | |
|---|-----------------------|-------------------------------|----------|-----------------|-----|----------|
| GO biological process complete | # | # | expected | Fold Enrichment | +/- | P value |
| histone H3-K4 methylation | 32 | 12 | 1.55 | 7.75 | + | 7.35E-04 |
| ↳ histone lysine methylation | 64 | 16 | 3.10 | 5.17 | + | 1.41E-03 |
| ↳ histone methylation | 84 | 19 | 4.07 | 4.67 | + | 4.79E-04 |
| ↳ histone modification | 337 | 43 | 16.31 | 2.64 | + | 1.67E-04 |
| ↳ covalent chromatin modification | 346 | 44 | 16.75 | 2.63 | + | 1.26E-04 |
| ↳ macromolecule metabolic process | 7359 | 496 | 356.16 | 1.39 | + | 1.25E-15 |
| ↳ organic substance metabolic process | 9032 | 524 | 437.13 | 1.20 | + | 2.08E-04 |
| ↳ metabolic process | 9443 | 535 | 457.02 | 1.17 | + | 4.42E-03 |
| ↳ chromatin organization | 636 | 80 | 30.78 | 2.60 | + | 2.59E-10 |
| ↳ chromosome organization | 984 | 93 | 47.62 | 1.95 | + | 1.08E-05 |
| ↳ organelle organization | 3133 | 214 | 151.63 | 1.41 | + | 8.13E-04 |
| ↳ cellular component organization | 5133 | 319 | 248.43 | 1.28 | + | 2.57E-03 |
| ↳ cellular process | 14559 | 773 | 704.62 | 1.10 | + | 9.14E-03 |
| ↳ cellular component organization or biogenesis | 5288 | 329 | 255.93 | 1.29 | + | 1.31E-03 |
| ↳ macromolecular complex subunit organization | 1983 | 148 | 95.97 | 1.54 | + | 9.07E-04 |
| ↳ primary metabolic process | 8601 | 519 | 416.27 | 1.25 | + | 4.01E-07 |
| ↳ cellular macromolecule metabolic process | 6693 | 475 | 323.93 | 1.47 | + | 3.28E-19 |
| ↳ cellular metabolic process | 8525 | 505 | 412.59 | 1.22 | + | 2.32E-05 |

Published History: Gene-Repeat overlap, entire genome

Yay! But, a wee challenge

We have exon names and counts

We really want genes (or transcripts) and counts
across the whole gene (or transcript)

What we have: Computer generated Exon IDs

uc002zmb.3_cds_0_0_chr22_17119391_r

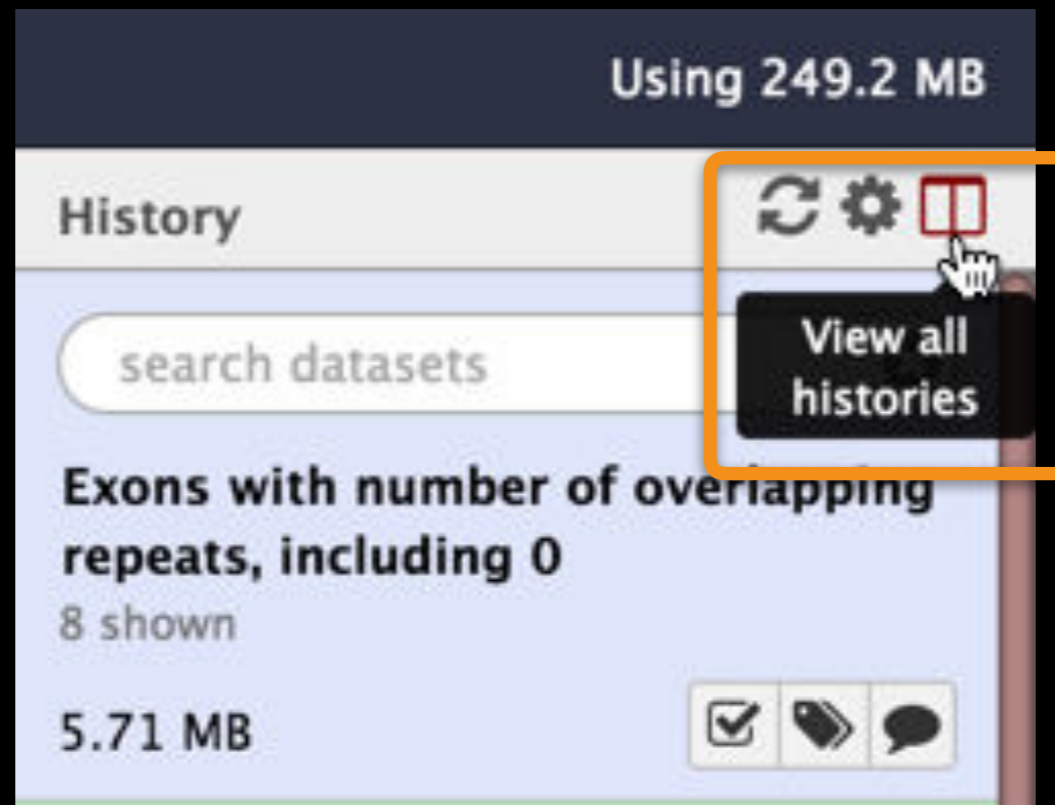
Transcript ID is embedded in Exon ID.*

How can we extract the Transcript ID from the Exon ID?

(With the transcript ID we can summarize counts for each transcript and/or get the gene ID.)

* How do we know that's a transcript ID?

Create another copy of your original history



Create another copy of your original history

The screenshot displays the UCSC Genome Browser interface with three history panels. The 'Current History' panel on the left lists several datasets, with the first one, 'Exons with number of overlapping repeats, including 0', highlighted. The middle panel shows a detailed view of a dataset, 'Genes with overlapping repeats', with 10 shown and 3.91 MB. The right panel shows a list of datasets, with the 'Copy' button highlighted in the top right corner. The 'Copy' button is located in the top right corner of the right panel, next to the 'Delete' and 'Purge' buttons. The 'Copy' button is highlighted with an orange box.

Done search histories search all datasets Create new

Current History

Exons with number of overlapping repeats, including 0
8 shown
5.71 MB

search datasets

Drag datasets here to copy them to the current history

8: # of overlapping repeats per exon, distribution

7: Exons and number of overlapping repeats.
14,875 lines
format: tabular, database: hg38

1 2
uc002zly.5_cds_10_0_chr22_17105853_f 1

6: Cut on data 5

5: Compare two Datasets on data 4 and data 1
14,083 regions
format: bed, database: hg38

join (GNU coreutils) 8.21
Copyright (C) 2013 Free Software Foundation, Inc

Genes with overlapping repeats
10 shown
3.91 MB

search datasets

10: Genes with # of overlapping repeats
228 lines
format: tabular, database: hg38
--Group by c5: max[c2]

1 2
AC007326.1 4

9: Join two Datasets on data 8 and data 6
623 lines
format: tabular, database: hg38

1 2 3 4 5
uc002zly.5 2 ENSG00000177663 ENST00000319363 IL

8: mart export.txt

Exons with overlapping
4 shown
3.79 MB

search datasets

4: Exons with overlapping repeats.

3: Join on data 2 and data 1

2: Repeats, chr22

1: Exons, chr22

Copy
Delete
Purge

Put the word Gene in the history name

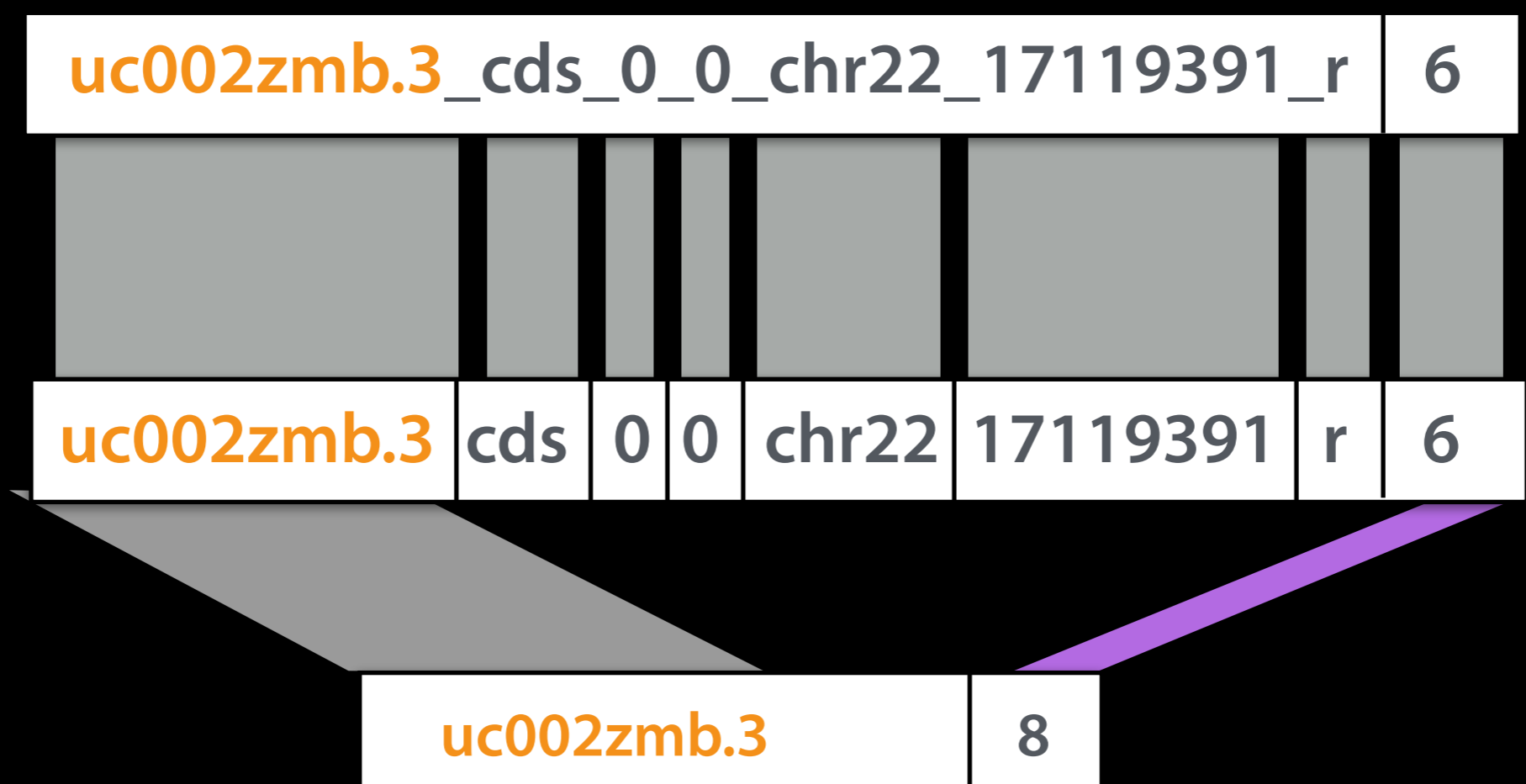
Extract the transcript ID

Split the exon ID into its constituent parts.

| | | | | | | | |
|-------------------------------------|-----|---|---|-------|----------|---|---|
| uc002zmb.3_cds_0_0_chr22_17119391_r | | | | | | | 6 |
| | | | | | | | |
| uc002zmb.3 | cds | 0 | 0 | chr22 | 17119391 | r | 6 |

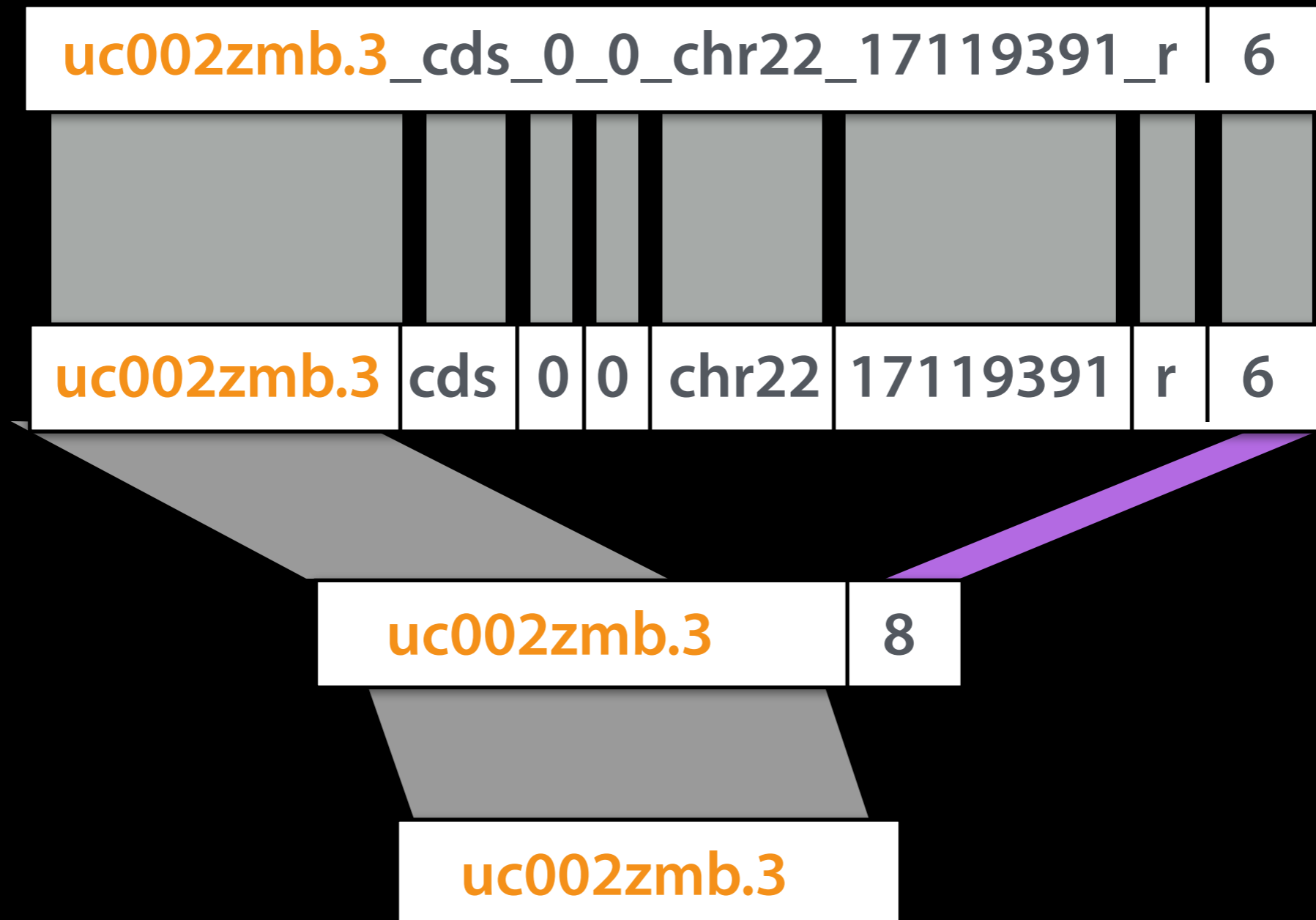
Text Manipulation → Convert delimiters to TAB
(convert underscores to tabs)

Sum the scores for all exons in each transcript



Join, Subtract and Group →
Group: by Transcript ID; Sum score

Get list of transcript IDs



Text Manipulation → Cut

Published History: Transcripts with # of overlapping repeats

Have Transcripts, now get Gene IDs

Save list of
Transcript IDs to
a file.

We'll upload it to
Ensembl BioMart

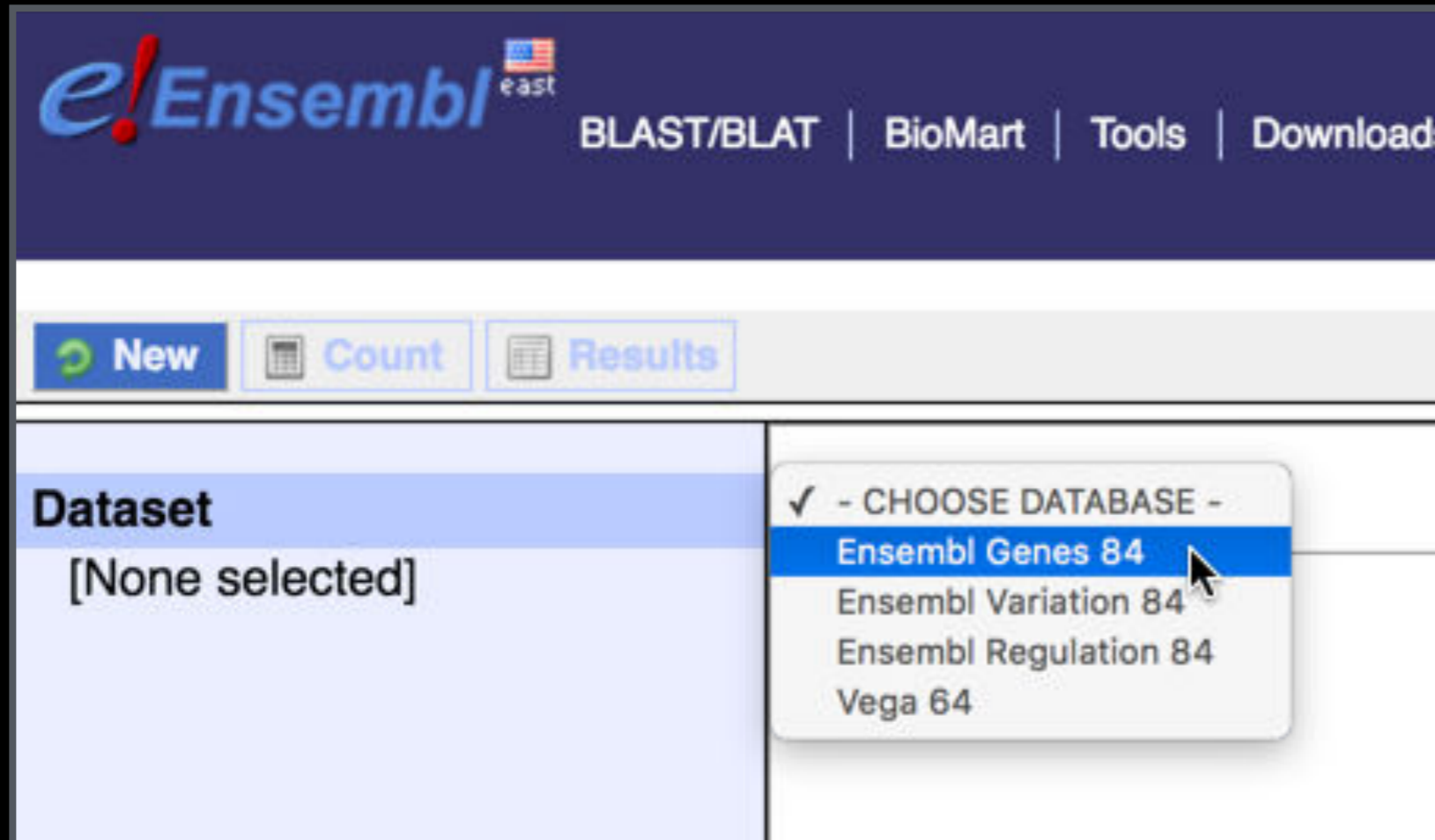
The screenshot shows the Ensembl BioMart interface. At the top, a query is titled "7: Transcripts with overlapping repeats" with a count of "628 lines". The format is set to "tabular" and the database is "hg38". Below the query title, there is a toolbar with icons for viewing, editing, deleting, and saving. The "save" icon, which is a floppy disk, is highlighted with an orange square. Below the toolbar, a table shows the first result: a blue bar with the number "1" and the transcript ID "uc002zly.5". At the bottom, another query is partially visible: "6: Transcripts with # of overlapping repeats".

Published History: Transcripts with # of overlapping repeats

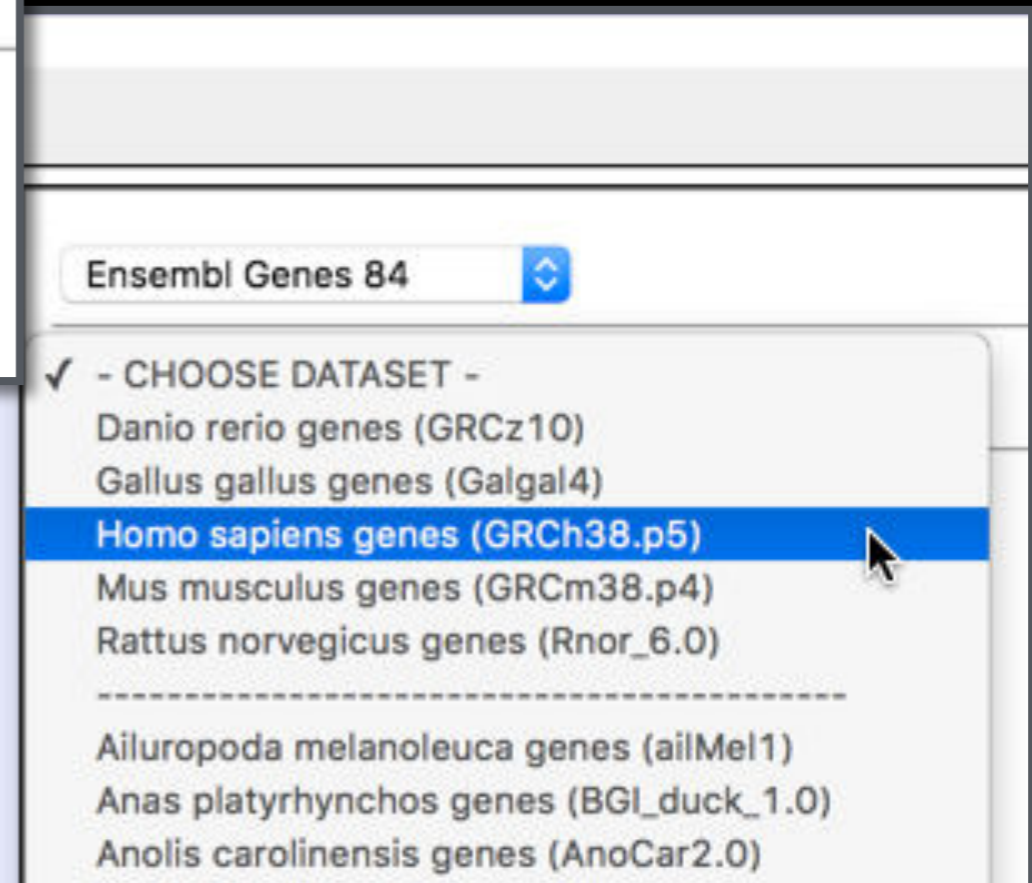
Ensembl BioMart

www.ensembl.org/biomart/martview

Specify Ensembl Genes 84, GRCh38.p5



The screenshot shows the Ensembl BioMart interface. At the top, the Ensembl logo is followed by navigation links: BLAST/BLAT, BioMart, Tools, and Downloads. Below this is a toolbar with three buttons: 'New' (with a green circular arrow icon), 'Count' (with a calculator icon), and 'Results' (with a table icon). The main content area has a 'Dataset' section on the left with the text '[None selected]'. To the right of this is a dropdown menu titled '- CHOOSE DATABASE -'. The menu is open, showing a list of options: 'Ensembl Genes 84' (highlighted in blue), 'Ensembl Variation 84', 'Ensembl Regulation 84', and 'Vega 64'. A mouse cursor is pointing at the 'Ensembl Genes 84' option.



This screenshot shows a continuation of the Ensembl BioMart interface. The 'Dataset' dropdown menu is open, displaying a list of species and their corresponding genome builds. The options are: 'Danio rerio genes (GRCz10)', 'Gallus gallus genes (Gallgal4)', 'Homo sapiens genes (GRCh38.p5)' (highlighted in blue), 'Mus musculus genes (GRCm38.p4)', and 'Rattus norvegicus genes (Rnor_6.0)'. Below these, there is a horizontal line and then 'Ailuropoda melanoleuca genes (ailMel1)', 'Anas platyrhynchos genes (BGI_duck_1.0)', and 'Anolis carolinensis genes (AnoCar2.0)'. A mouse cursor is pointing at the 'Homo sapiens genes (GRCh38.p5)' option.

Ensembl BioMart:

New **Count** **Results** **★ URL** **XML** **Perl** **Help**

Dataset
Homo sapiens genes (GRCh38.p5)

Filters
[None selected]

Attributes
Ensembl Gene ID
Ensembl Transcript ID
Associated Gene Name

Dataset
[None Selected]

☒ **Features** ☐ **Variant (Germline)**
☐ **Structures** ☐ **Variant (Somatic)**
☐ **Homologues** ☐ **Sequences**

☐ **GENE:**

Ensembl

- ☒ Ensembl Gene ID
- ☒ Ensembl Transcript ID
- ☐ Ensembl Protein ID
- ☐ Ensembl Exon ID
- ☐ Description
- ☐ Chromosome Name
- ☐ Gene Start (bp)
- ☐ Gene End (bp)
- ☐ Strand
- ☐ Band
- ☐ Transcript Start (bp)
- ☐ Transcript End (bp)
- ☐ Transcription Start Site (TSS)
- ☐ Transcript length (including UTRs and CDS)
- ☐ Transcript Support Level (TSL)
- ☐ GENCODE basic annotation

Phenotype

- ☐ APPRIS annotation
- ☒ Associated Gene Name
- ☐ Associated Gene Source
- ☐ Associated Transcript Name
- ☐ Associated Transcript Source
- ☐ Transcript count
- ☐ % GC content
- ☐ Gene type
- ☐ Transcript type
- ☐ Source (gene)
- ☐ Source (transcript)
- ☐ Status (gene)
- ☐ Status (transcript)
- ☐ Version (gene)
- ☐ Version (transcript)

Specify attributes to put in output report

Ensembl BioMart:

New Count Results URL XML Perl Help

Dataset
Homo sapiens genes (GRCh38.p5)

Filters
[None selected]

Attributes
Ensembl Gene ID
Ensembl Transcript ID
Associated Gene Name
UCSC ID

Dataset
[None Selected]

☐ GOSlim GOA Accession(s) ☐ GOSlim GOA Description

External References (max 3)

☐ ArrayExpress ☐ MIM Gene Description
☐ ChEMBL ID(s) ☐ miRBase Accession(s)
☐ Clone based Ensembl gene name ☐ miRBase ID(s)
☐ Clone based Ensembl transcript name ☐ miRBase transcript name
☐ Clone based VEGA gene name ☐ PDB ID
☐ Clone based VEGA transcript name ☐ Protein (Genbank) ID [e.g. AAA0248]
☐ CCDS ID ☐ Reactome ID
☐ Database of Aberrant 3' Splice Sites (DBASS3) IDs ☐ Reactome gene ID
☐ DBASS3 Gene Name ☐ Reactome transcript ID
☐ Database of Aberrant 5' Splice Sites (DBASS5) IDs ☐ RefSeq mRNA [e.g. NM_001195597]
☐ DBASS5 Gene Name ☐ RefSeq mRNA predicted [e.g. XM_0]
☐ EMBL (Genbank) ID ☐ RefSeq ncRNA [e.g. NR_002834]
☐ Ensembl Human Transcript IDs ☐ RefSeq ncRNA predicted [e.g. XR_1]
☐ Ensembl Human Translation IDs ☐ RefSeq Protein ID [e.g. NP_001005]
☐ LRG to Ensembl link gene ☐ RefSeq Predicted Protein ID [e.g. XI]
☐ LRG to Ensembl link transcript ☐ Rfam ID
☐ EntrezGene ID ☐ Rfam transcript name
☐ EntrezGene transcript name ID ☒ UCSC ID
☐ Human Protein Atlas Antibody ID ☐ Unigene ID
☐ VEGA gene ID(s) (OTTG) ☐ UniParc
☐ VEGA transcript ID(s) (OTTT)

Specify attributes to put in output report

Ensembl BioMart:

The screenshot shows the Ensembl BioMart interface. On the left sidebar, the 'Results' tab is highlighted with an orange box and a red number 4. Below it, the 'Filters' tab is also highlighted with an orange box and a red number 1. The main panel is titled 'Please restrict your query using criteria below' and contains several filter sections. The 'GENE' section is expanded, showing three options: 'Limit to genes (external references)...', 'Input external references ID list [Max 500 advised]', and 'Limit to genes (microarray probes/probesets)...'. The 'Input external references ID list' option is selected with a blue checkmark. To its right, there is a dropdown menu for 'with HGNC ID(s)' with 'Only' selected, and a text input field containing 'UCSC ID(s) [e.g. uc002cqj.3]' which is highlighted with an orange box and a red number 2. Below this, there is a 'Choose File' button and a text input field containing 'Galaxy6-[Trans...eats].tabular' which is highlighted with an orange box and a red number 3. The 'Limit to genes (microarray probes/probesets)...' option is also visible, with a dropdown menu for 'with Affymetrix Microarray huex 1 0 st v2 probeset ID(s)' and 'Only' selected. At the bottom, there is a section for 'Input microarray probes/probesets ID list [Max 500 advised]' with a dropdown menu for 'Codelink probe ID(s) [e.g. GE550734]' and a 'Choose File' button.

New Count Results

URL XML Perl Help

Please restrict your query using criteria below
(If filter values are truncated in any lists, hover over the list item to see the full text)

REGION:

GENE:

☐ Limit to genes (external references)...

☒ Input external references ID list [Max 500 advised]

☐ Limit to genes (microarray probes/probesets)...

☐ Input microarray probes/probesets ID list [Max 500 advised]

with HGNC ID(s) Only Excluded

UCSC ID(s) [e.g. uc002cqj.3]

Choose File Galaxy6-[Trans...eats].tabular

with Affymetrix Microarray huex 1 0 st v2 probeset ID(s) Only Excluded

Codelink probe ID(s) [e.g. GE550734]

Choose File No file chosen

Specify which genes we want this information for

Ensembl BioMart:

[New](#) [Count](#) [Results](#)

[★ URL](#) [XML](#) [Perl](#) [Help](#)

Dataset
Homo sapiens genes (GRCh38.p5)

Filters
UCSC ID(s) [e.g. uc002cqj.3]:
[ID-list specified]

Attributes
Ensembl Gene ID
Ensembl Transcript ID
Associated Gene Name
UCSC ID

Dataset
[None Selected]

Export all results to
Email notification to

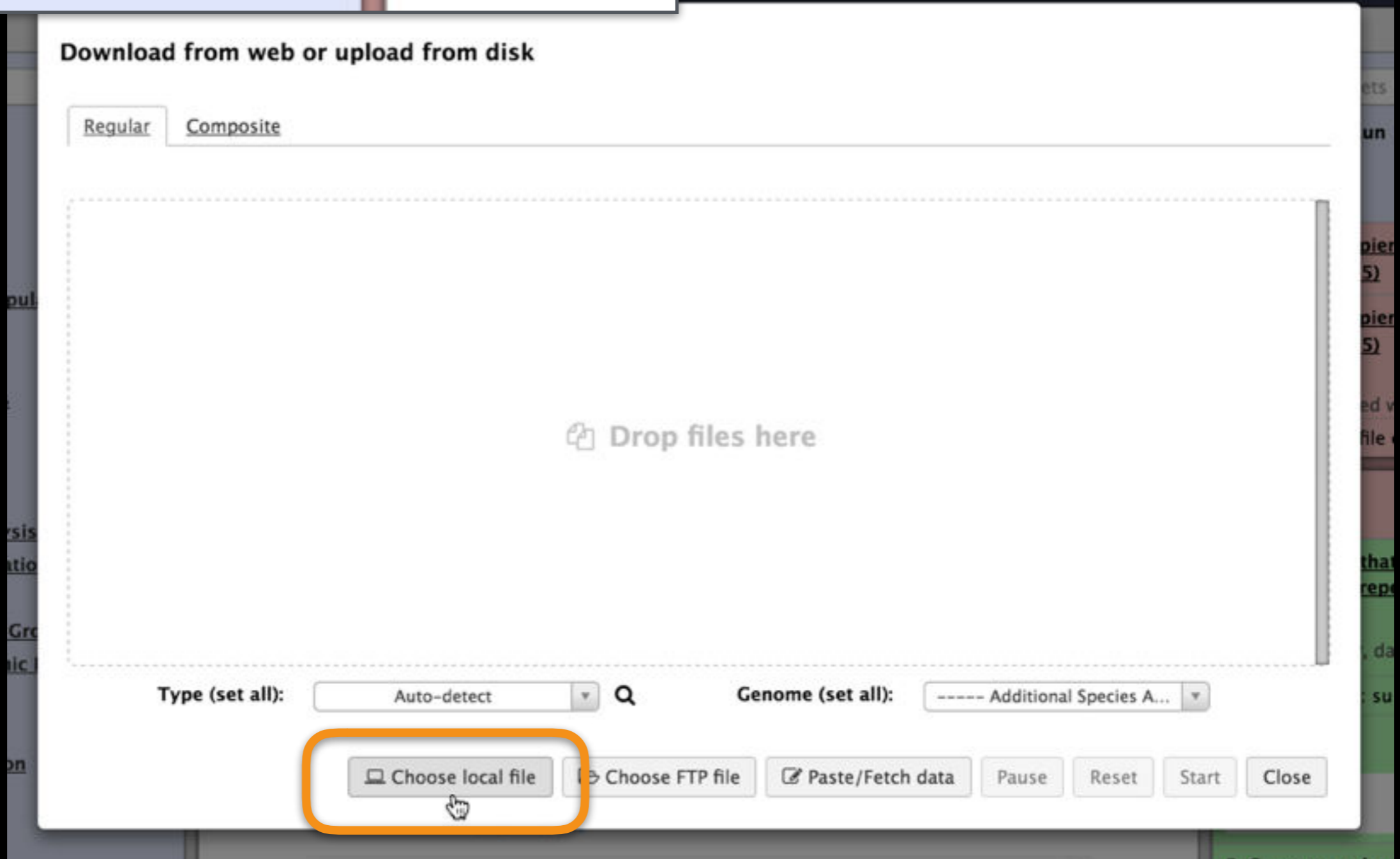
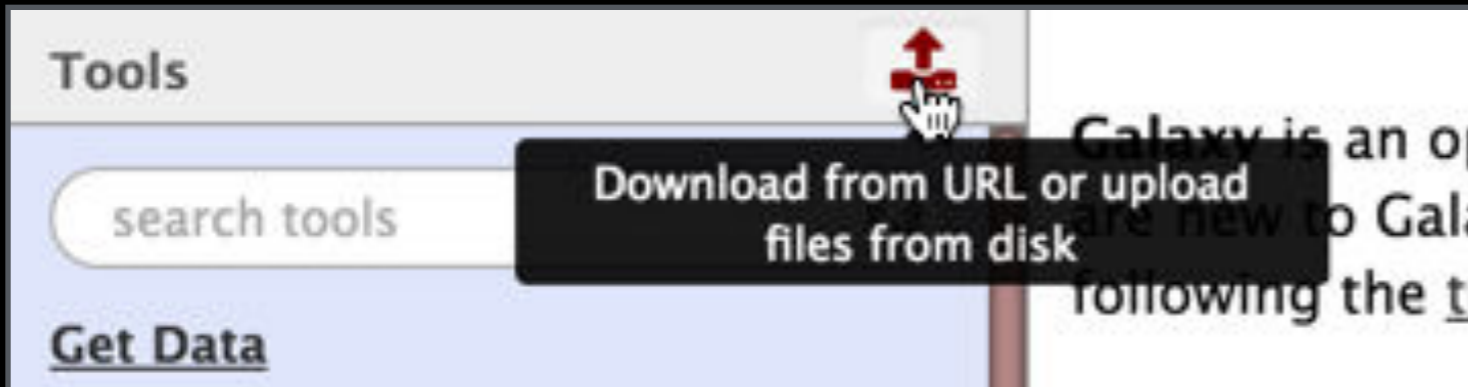
File TSV ☐ Unique results only [Go](#)

View 10 rows as HTML ☐ Unique results only

| Ensembl Gene ID | Ensembl Transcript ID | Associated Gene Name | UCSC ID |
|---------------------------------|---------------------------------|-------------------------|----------------------------|
| ENSG00000177663 | ENST00000319363 | IL17RA | uc002zly.5 |
| ENSG00000183307 | ENST00000331437 | CECR6 | uc002zmb.3 |
| ENSG00000099968 | ENST00000317582 | BCL2L13 | uc002zmw.5 |
| ENSG00000099968 | ENST00000543133 | BCL2L13 | uc002zmx.4 |
| ENSG00000099968 | ENST00000355028 | BCL2L13 | uc002zmy.5 |
| ENSG00000099968 | ENST00000418951 | BCL2L13 | uc002zmz.4 |
| ENSG00000243156 | ENST00000441493 | MICAL3 | uc002zng.5 |
| ENSG00000184979 | ENST00000215794 | USP18 | uc002zny.4 |
| ENSG00000100056 | ENST00000252137 | DGCR14 | uc002zou.4 |
| ENSG00000100075 | ENST00000451283 | SLC25A1 | uc002zoy.5 |


Save the results to a file for uploading into Galaxy

Get Genes into Galaxy



Chose local file, then Start, then Close






Get Gene IDs into Galaxy

8: mart_export.txt   



624 lines

format: **tabular**, database: ?

uploaded tabular file







    

| 1 | 2 |
|-----------------|-----------------------|
| Ensembl Gene ID | Ensembl Transcript ID |

7: Transcripts with overlapping repeats   

628 lines

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

| |
|------------|
| 1 |
| uc002zly.5 |

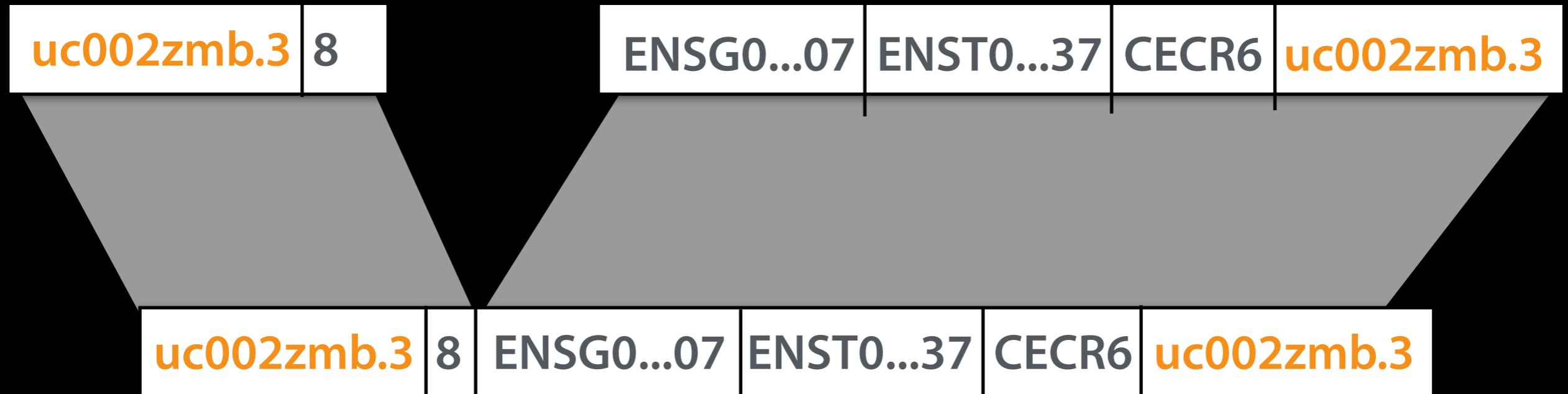
Upload file from BioMart.
Note that we lost 4-5 transcripts

Do we care?
Can we find out which were lost?

Unite our Transcript Scores with Biomart info

Transcript Scores

Biomart Info






Join, Subtract and Group →
Join: Transcripts with score and Biomart dataset;
join on UCSC transcript ID

Unite our Transcript Scores with Biomart info

Join two Datasets side by side on a specified field (Galaxy Version 2.0.2) Options




Join

   8: mart_export.txt

using column

Column: 4

with

   6: Transcripts with # overlapping repeats

and column

Column: 1

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


No

Keep lines of first input that are incomplete

No

Fill empty columns

No

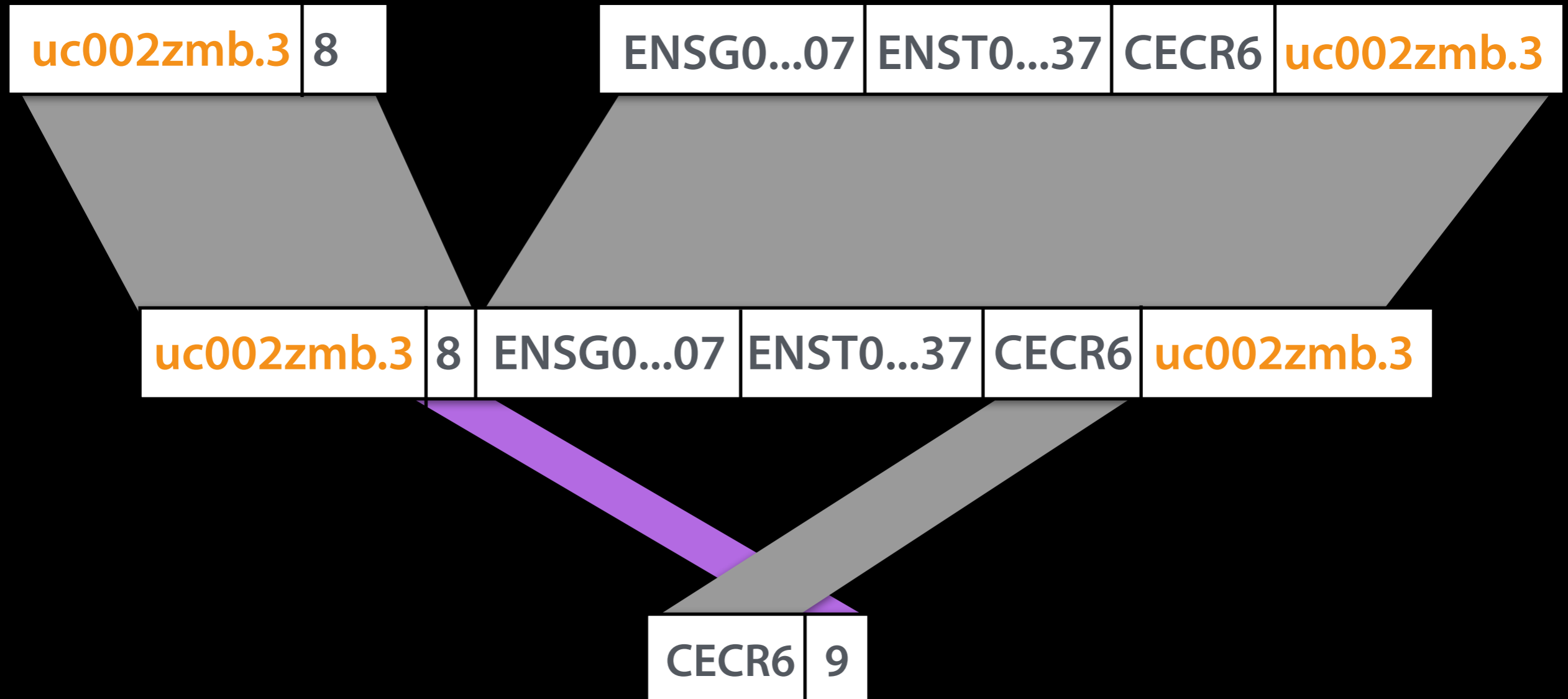
 **Execute**

Join, Subtract and Group → Join

Assign scores to genes

Transcript Scores

Biomart Info







Join, Subtract and Group →
Group: by gene symbol; Max score




Published History: Genes with overlapping repeats








Now have a list of genes with # overlapping repeats

| 1 | 2 |
|------------|---|
| AC007326.1 | 4 |
| ACR | 2 |
| ADM2 | 1 |
| ADRBK2 | 1 |
| ADSL | 2 |
| ANKRD54 | 1 |
| AP000349.2 | 1 |
| APOBEC3B | 2 |
| APOBEC3F | 1 |
| APOBEC3H | 1 |
| APOL3 | 1 |
| APOL4 | 2 |
| APOL5 | 1 |
| APOL6 | 1 |
| ARFGAP3 | 1 |
| ARHGAP8 | 2 |
| ARSA | 1 |
| ASCC2 | 2 |
| ASPHD2 | 1 |
| ATP6V1E1 | 1 |








History   

search datasets 

Genes with overlapping repeats
10 shown
3.91 MB   

10: Genes with # of overlapping repeats   
228 lines
format: tabular, database: hg38
--Group by c5: max[c2]
     

| 1 | 2 |
|------------|---|
| AC007326.1 | 4 |

9: Join two Datasets on data 8 and data 6   
623 lines
format: tabular, database: hg38
     

Published History: Genes with overlapping repeats