

# Introduction to Galaxy

---

University of Utah  
November 16, 2016

Dave Clements

Galaxy Team

Johns Hopkins University

<http://galaxyproject.org/>



#usegalaxy @galaxyproject

# 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
- 5:00 Done

# 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](http://usegalaxy.org)


The screenshot displays the Galaxy web interface. On the left is a navigation menu with categories like 'Get Data', 'Text Manipulation', and 'NGS: DeepTools'. The main content area features a text block about Galaxy's open-source nature and a 'GAME 2017 Melbourne 3-9 February' event announcement with a red globe logo and the text 'Talk abstracts due 30 November'. Below the event announcement are logos for Penn State, Johns Hopkins, TACC, and Cyverse. On the right, a 'Tweets by @galaxyproject' section shows two tweets: one retweeted by Pratik Jagtap and another from the Galaxy Project about a workshop in Bordeaux.

**Galaxy** Analyze Data Workflow Shared Data Visualization Help User Using 0%

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](#)  
[NGS: Assembly](#)  
[Operate on Genomic Intervals](#)  
[Statistics](#)  
[Graph/Display Data](#)  
[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](#).


 **GAME 2017**  
Melbourne  
3-9 February

Talk abstracts due 30 November

PENNSSTATE  
JOHNS HOPKINS  
TACC  
CYVERSE

Tweets by @galaxyproject

Galaxy Project Retweeted

 **Pratik Jagtap** @pratikomics  
z.umn.edu/hayahym 'How are you - and How's your Microbiome?' an article in @tAnaSci Issue#1116 #microbiome #metaproteomics #usegalaxyp

Galaxy Project @galaxyproject  
28-30 Nov: Analyse avancée de séquences, Bordeaux  
cnrsformation.cnrs.fr/stage-16148-An...  
#usegalaxy @cgfbordeaux

**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 ability to display, compare, store, and analyze the results of microbiome analysis. The portal is a community-driven platform for sharing and analyzing microbiome data. It includes a variety of tools for data management, analysis, and visualization, as well as a user-friendly interface for data exploration and reporting.




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



(GIGA)<sup>n</sup> Galaxy  
by CBIIT

Integrated publishing of workflows from GIGA<sup>n</sup> 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](http://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

# Galaxy on the Cloud: CloudLaunch

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

- Directly launch a Galaxy instance on AWS or Jetstream
- Uses CloudMan

## Galaxy Cloud Launch

Easily launch your own cloud servers for use with Galaxy and CloudMan. See [this page](#) for detailed instructions on how to get started.

Cloud

-----  
Amazon - Tokyo (AWS EC2)  
✓ Amazon - Virginia (AWS EC2)  
Amazon - Ireland (AWS EC2)  
Jetstream (development) (OpenStack)

Provide details below that must match (ie, exist on) the chosen cloud.

Access key

Your cloud account API access key. For the Amazon cloud, available from the [security credentials page](#).

Secret key

Your cloud account API secret key. For the Amazon cloud, also available from the [security credentials page](#).

Cluster name

or

Name of your cluster used for identification and restarting. If creating a new cluster, type any name you like.

Password

Your choice of password, for the CloudMan web interface and accessing the server via ssh.

Instance type

Compute Optimized large (c3.large) (2 vCPU / 3.75GB R ▾)

Type (ie, virtual hardware configuration) of the server to start.

# Galaxy on the Cloud: Jetstream

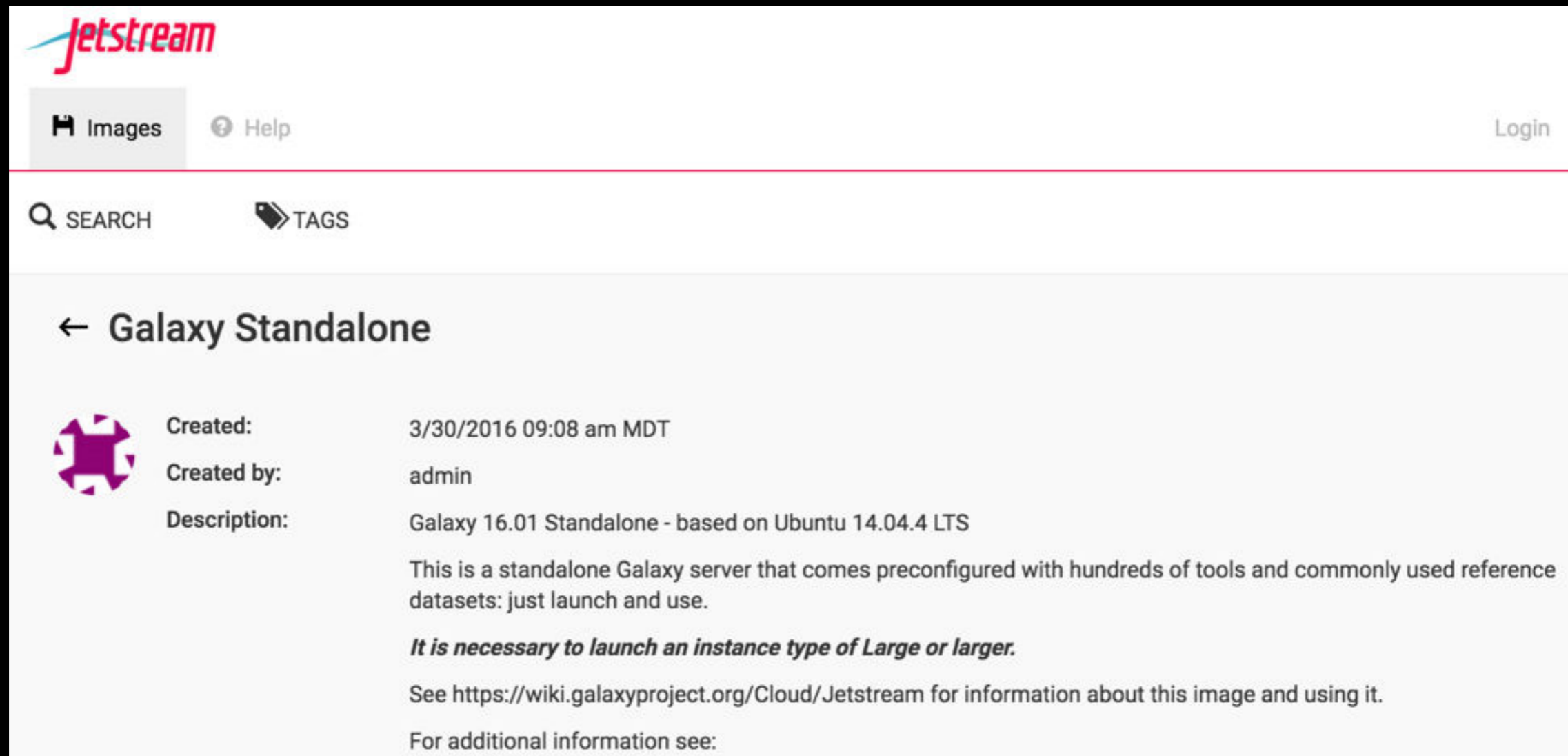
<https://wiki.galaxyproject.org/Cloud/Jetstream>

**Jetstream**

Images Help Login

SEARCH TAGS

← Galaxy Standalone

 Created: 3/30/2016 09:08 am MDT  
Created by: admin  
Description: Galaxy 16.01 Standalone - based on Ubuntu 14.04.4 LTS

This is a standalone Galaxy server that comes preconfigured with hundreds of tools and commonly used reference datasets: just launch and use.

*It is necessary to launch an instance type of Large or larger.*

See <https://wiki.galaxyproject.org/Cloud/Jetstream> for information about this image and using it.

For additional information see:

US based researchers can request an XSEDE allocation and then run Galaxy on Jetstream

U XSEDE Champion: Anita Orendt

# 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

5:00 Done

# 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

[bit.ly/utah\\_cloud1](http://bit.ly/utah_cloud1) (54.224.69.148)

[bit.ly/utah\\_cloud2](http://bit.ly/utah_cloud2) (54.83.150.22)

[bit.ly/utah\\_cloud3](http://bit.ly/utah_cloud3) (54.144.233.255)

# 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 intersections, to retrieve 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 [Table Browser](#). To examine the biological function of your set through annotation enrichments, send the data to [Table Browser](#) for use with diverse computational tools. Refer to the [Credits](#) page for the list of contributors and for these data. All tables can be downloaded in their entirety from the [Sequence and Annotation Download](#)

**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=

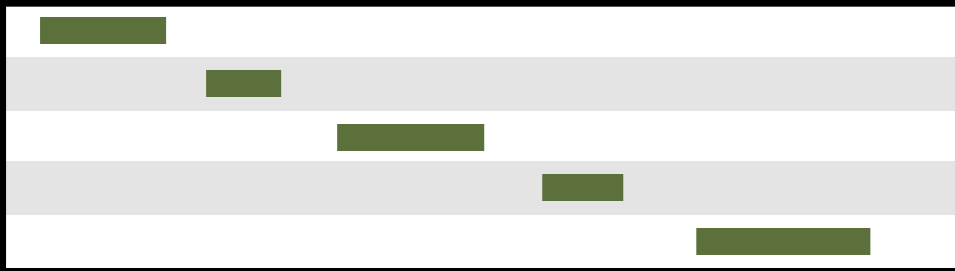
visibility=

url=

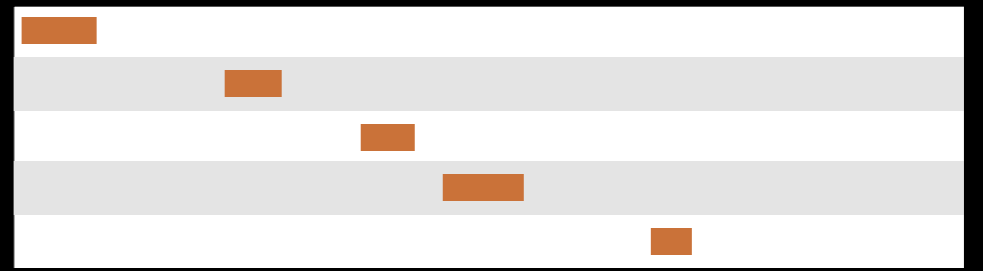
### Create one BED record per:

- 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.

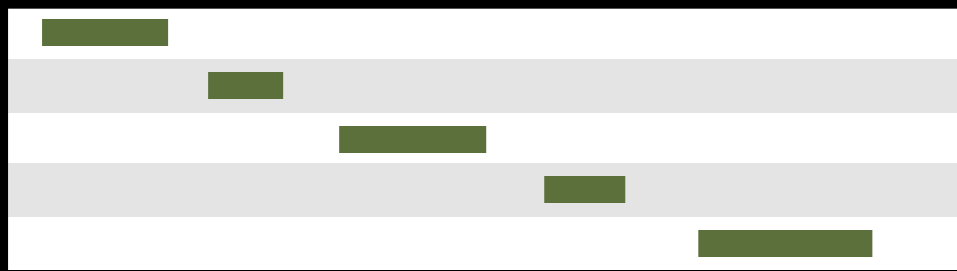


Exons

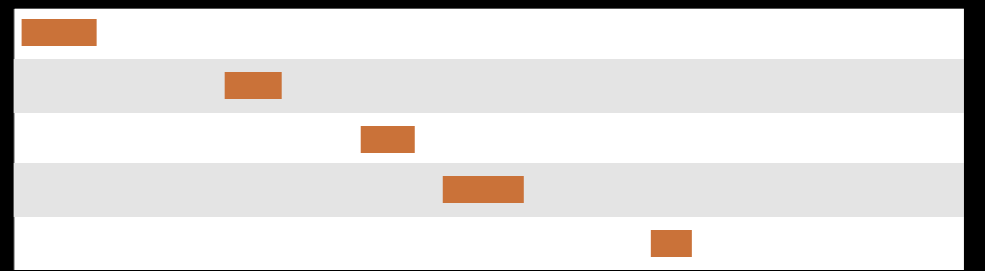


Repeats

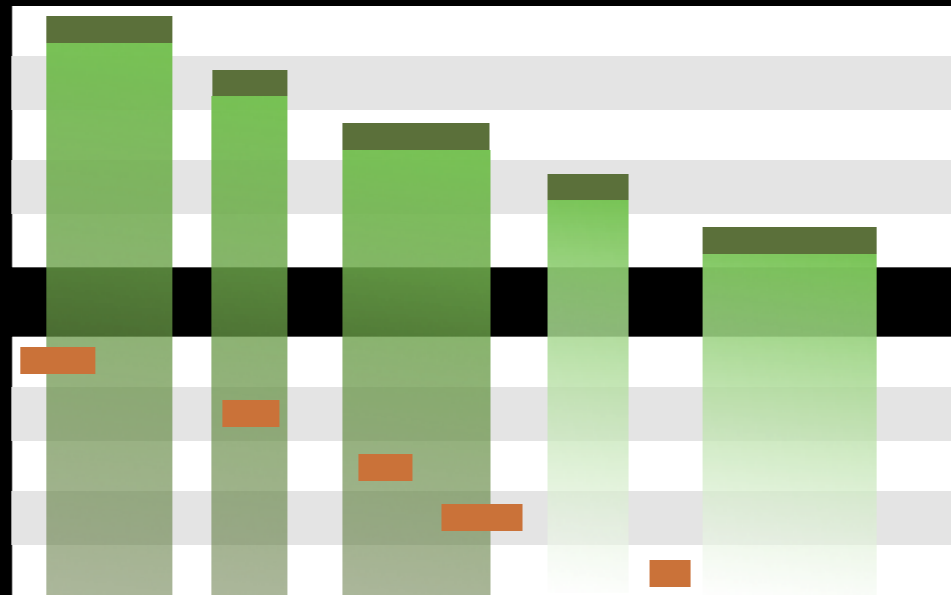
(Identify which exons have Repeats)



Exons



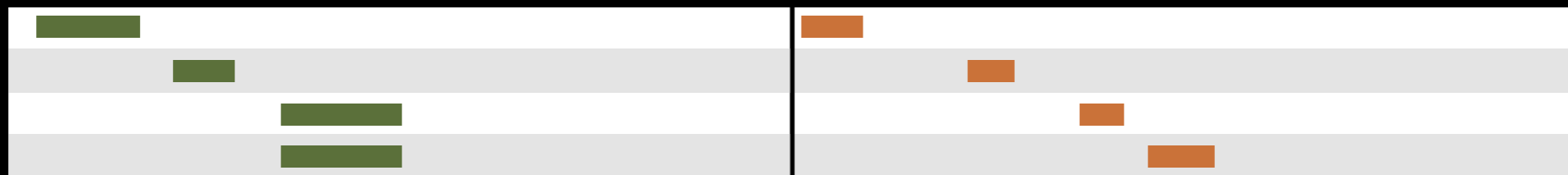
Repeats



Exons

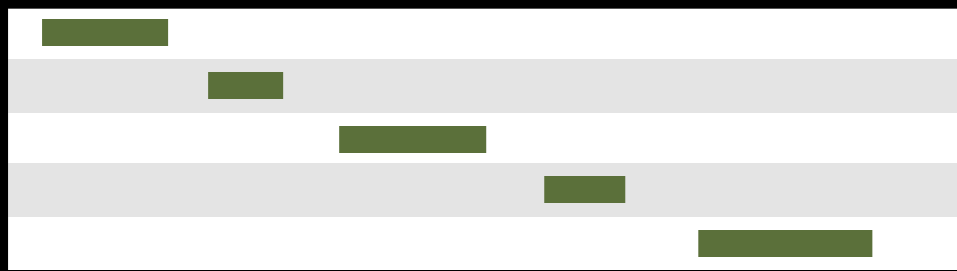
Repeats

Overlap pairings

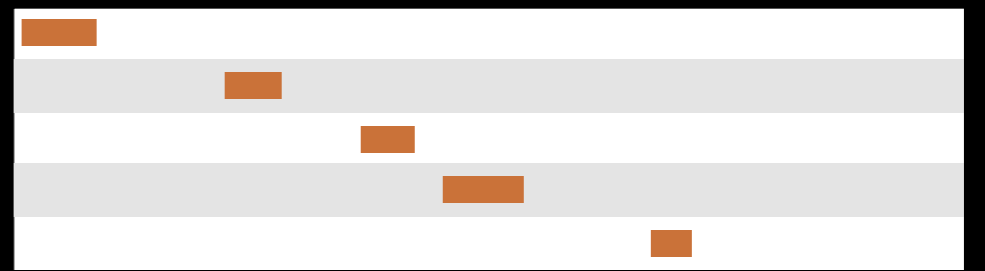


Operate on Genomic Intervals → Join  
 (Identify which exons have Repeats)

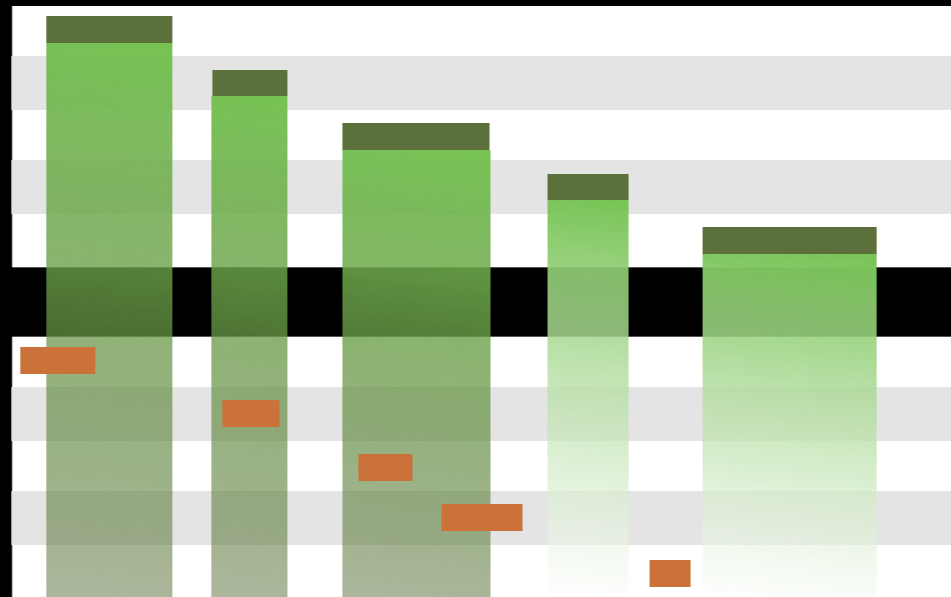




Exons



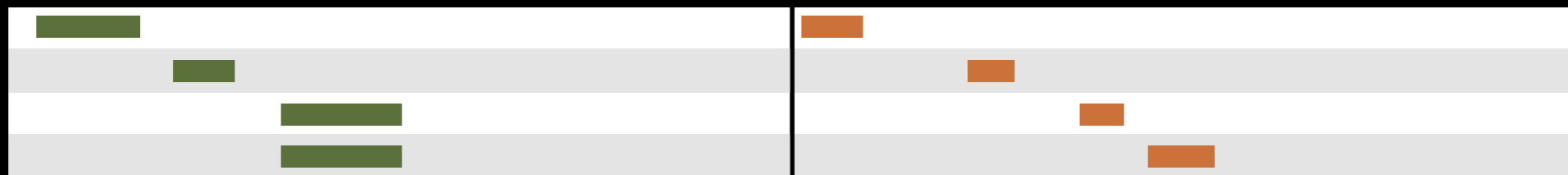
Repeats



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.

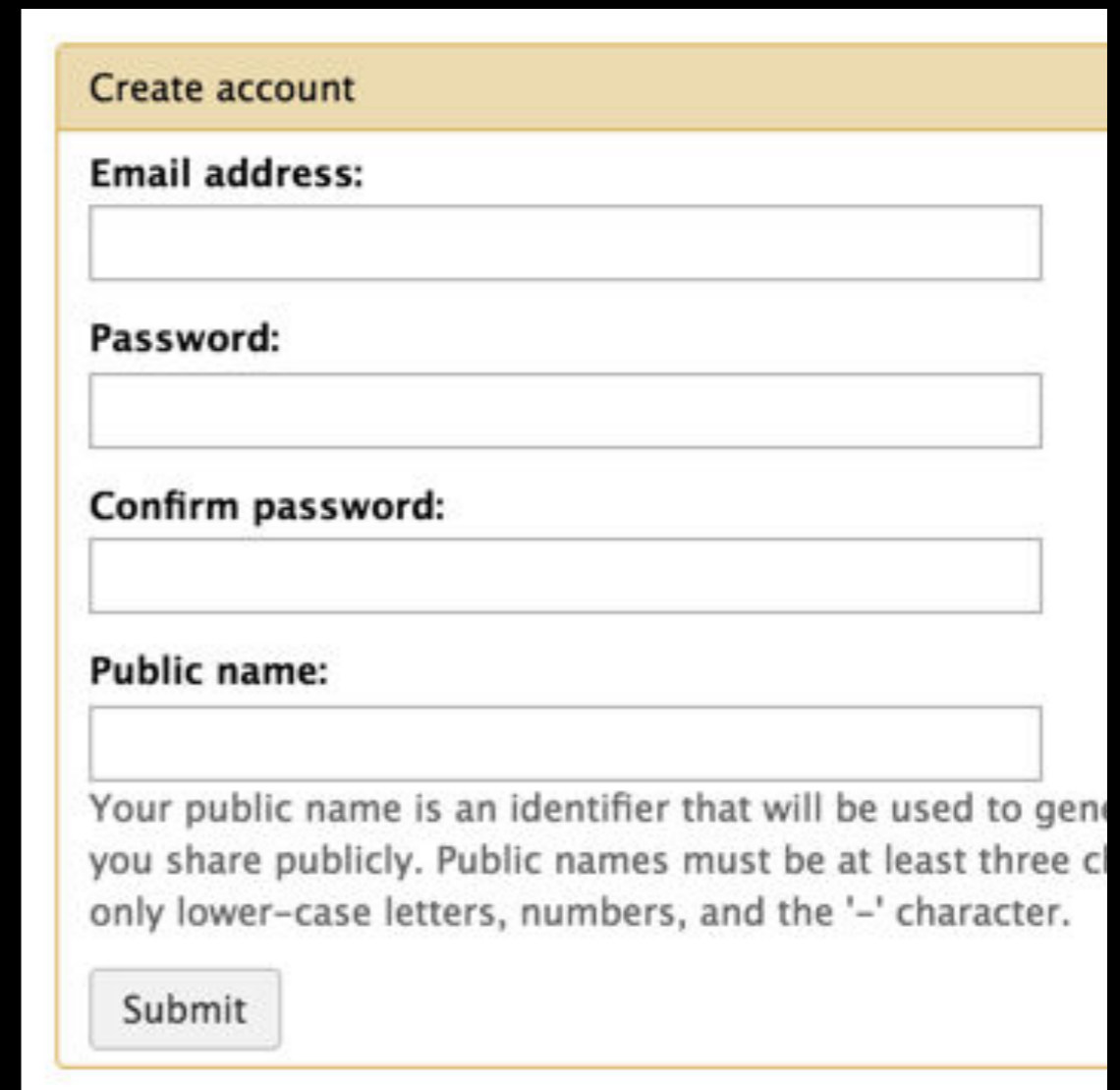
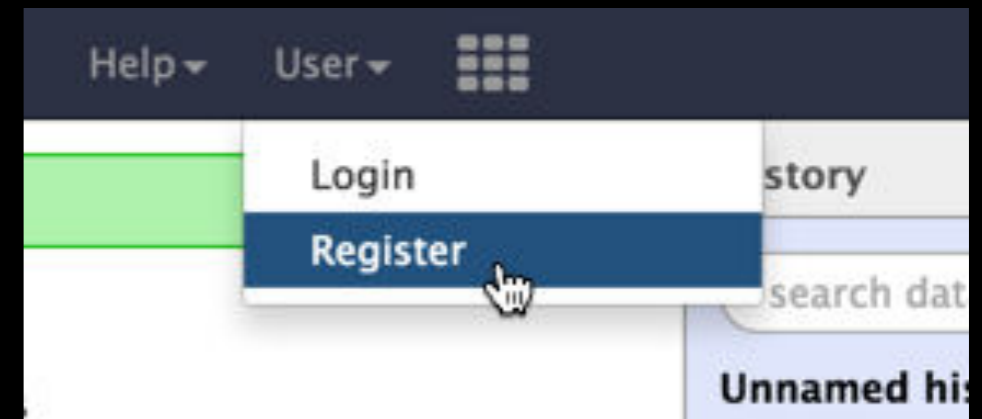
# But first, create a login

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

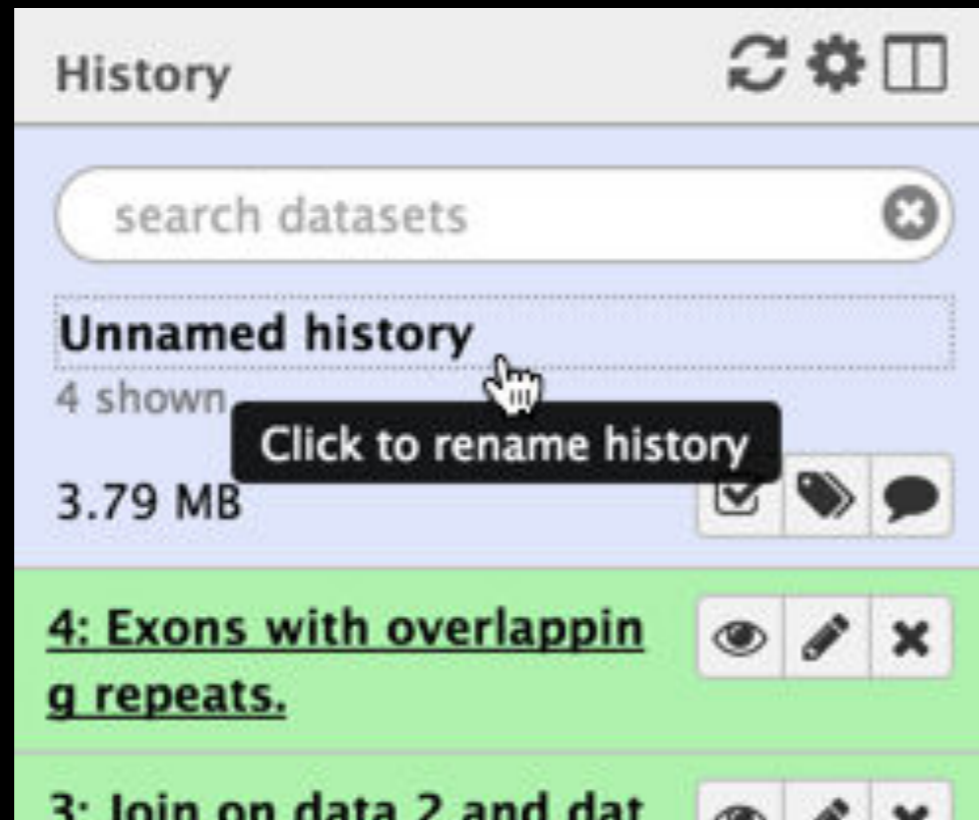
Use an email address you can remember.

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 with the text 'Create account'. Below the header are four input fields: 'Email address:', 'Password:', 'Confirm password:', and 'Public name:'. Below the 'Public name:' field is a paragraph of text: 'Your public name is an identifier that will be used to generate your public profile. Public names must be at least three characters long and can only contain lower-case letters, numbers, and the '-' character.' At the bottom of the form is a 'Submit' button.

# 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.

The screenshot shows a context menu with the following items:

- HISTORY LISTS
  - Saved Histories
  - Histories Shared with Me
- HISTORY ACTIONS
  - Create New
  - Copy History** (highlighted with an orange box and a mouse cursor)
  - 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

# Exons & Repeats: Pick an Exercise

1. Report the number of overlapping repeats that each exon has (what we just did), but also **include exons with 0 overlapping repeats in the output.**
2. Create the **list of exons** with overlapping repeats, **in 6-column BED format.** Set the score column to be the number of overlapping repeats that exon has.

Everything you need will be in these toolboxes

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



# 1. All exons, even those with 0 overlaps

Can take advantage of fact that score column of all exons is 0 to begin with.

**Join, subtract and group** is a good place to start.

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

## 2. List of exons with overlaps, in BED

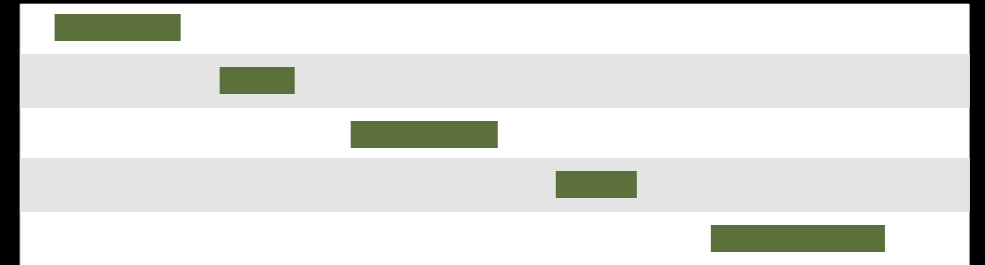
Can be done in two steps, one of them a Cut,  
plus an edit attributes step at the end:

The screenshot shows the Galaxy web interface. On the left, the 'Datatype' step is active, with a dropdown menu set to 'bed'. Below the dropdown, a 'Save' button is visible. On the right, the 'History' panel shows a dataset titled '6: Exons with overlapping repeats, in BED' with 792 regions and a format of 'interval'. The 'edit' icon for this dataset is highlighted with an orange box. Below the history entry, a table header is visible with columns: 1. Chrom, 2. Start, 3. End, 4. Name.


Published History: Exons with overlapping repeats, in BED

	1
	1
	2

Exon overlap counts



Exons

	1		0
	1		0
	2		0

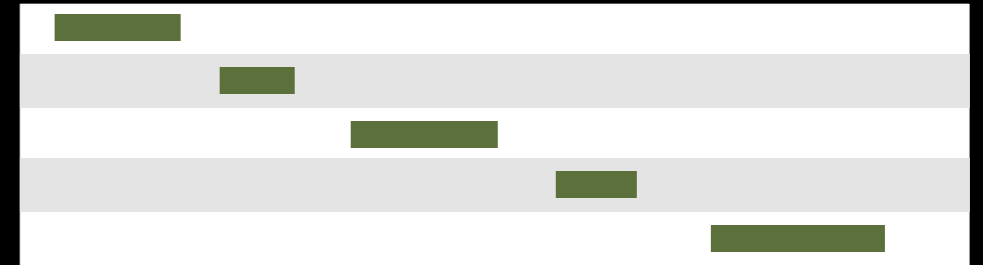
Join on exon name

Join, Subtract, and Group → Join

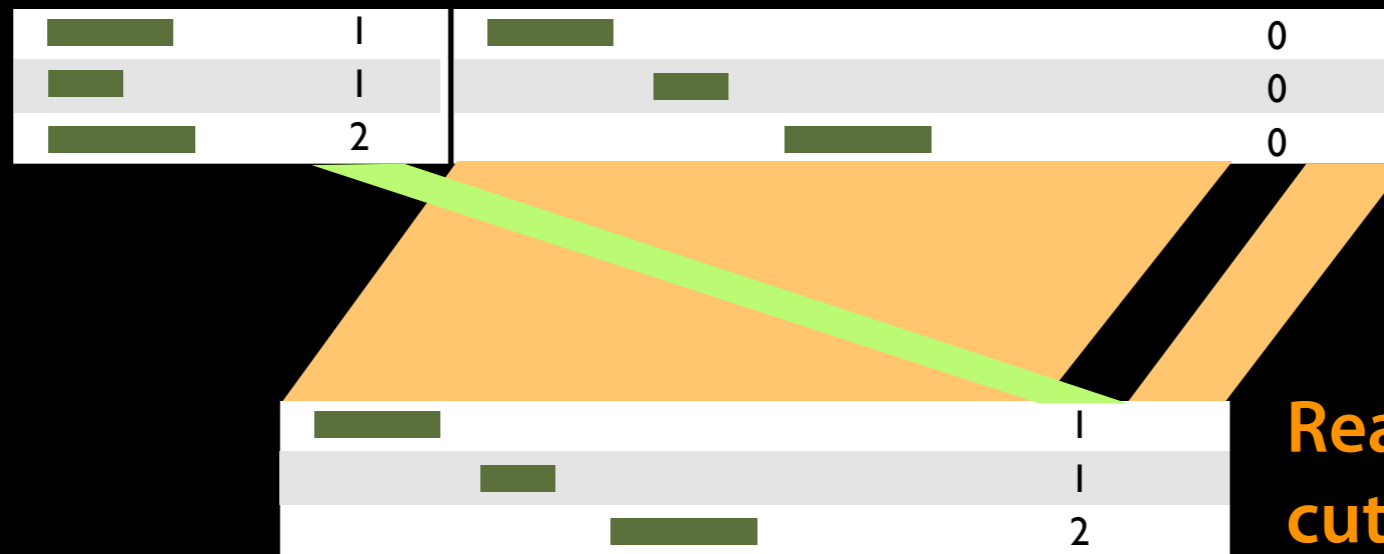
(Incorporate the overlap count with rest of Exon information)

█	1
█	1
█	2

Exon overlap counts



Exons



Join on exon name

Rearrange columns w/  
cut

Text Manipulation → Cut

(Incorporate the overlap count with rest of Exon information)

# 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

5:00 Done



[gmod.org](http://gmod.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

5:00 Done



# 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:**

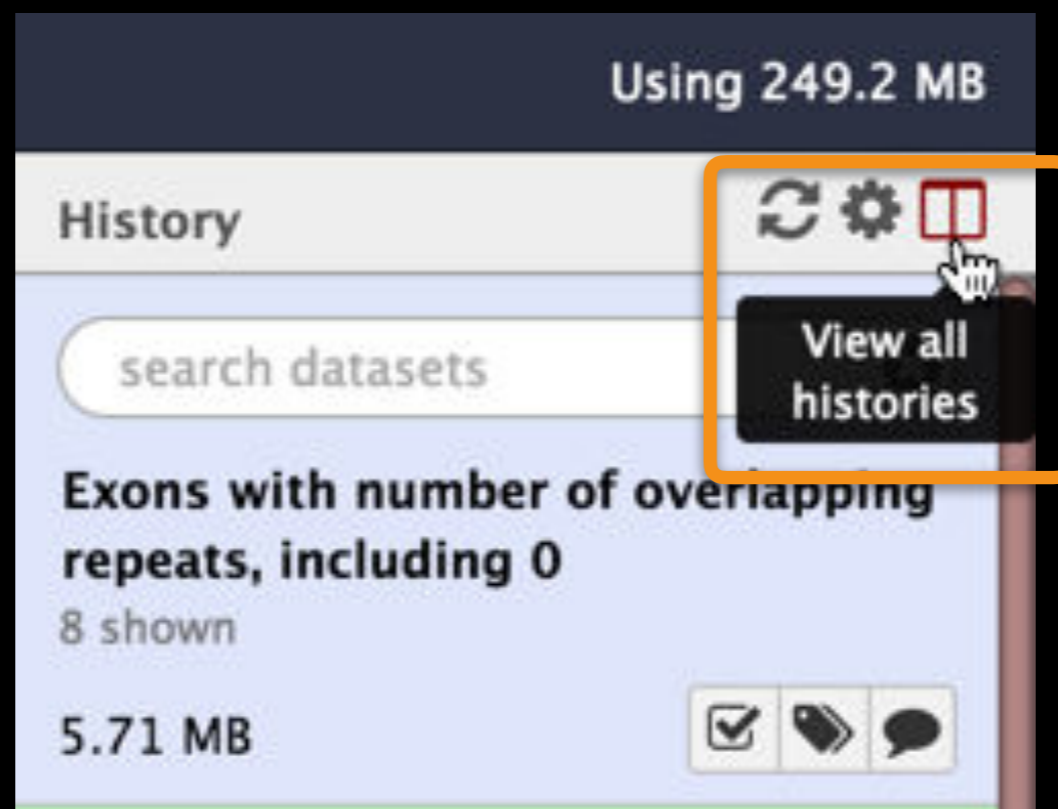
A series of analysis steps

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



# Get back to the original history

The screenshot displays the Galaxy web interface with three history panels. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The top right corner shows 'Using 2.1 MB'. The interface is divided into three main sections, each with a 'Switch to' button highlighted in orange and numbered '1'. The leftmost section is titled 'Current History' and contains a 'Done' button highlighted in orange and numbered '2'. Below the 'Done' button is a search bar for 'histories'. The middle section is titled 'Exons with number of overlapping repeats, including 0' and contains a search bar for 'all datasets'. The rightmost section is titled 'Exon Repeat Counts, chr22' and contains a search bar for 'all datasets'. Each section displays a list of data analysis steps with their respective outputs and icons for viewing, editing, and deleting. The steps are numbered 1 through 7. The first step in each section is '1: Exons, chr22'. The second step is '2: Repeats, chr22'. The third step is '3: Join on data 2 and data 1'. The fourth step is '4: Exons with number of overlapping repeats'. The fifth step is '5: Compare two Datasets on data 4 and data 1'. The sixth step is '6: Cut on data 5'. The seventh step is '7: Exons with # of overlapping repeats, including those with 0 overlaps'. The output of the seventh step is a table with 2 columns and 1 row: 'uc002z1y.5\_cds\_10\_0\_chr22\_17105853\_f 1'. The interface also includes a 'Loading histories...' indicator on the right side.

# 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' window with a search bar and a list of history items. A context menu is open over the list, showing various actions categorized into sections:

- HISTORY LISTS**
  - Saved Histories
  - Histories Shared with Me
- HISTORY ACTIONS**
  - Create New
  - Copy History
  - Share or Publish
  - Show Structure
  - Extract Workflow** (highlighted)
  - 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



# 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

The screenshot displays a workflow editor interface with three main sections: Tools, Workflow Canvas, and Details.

- Tools:** A sidebar on the left containing a search bar and a list of tool categories such as 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.
- Workflow Canvas:** The central workspace titled "Workflow Canvas | count overlapping features" shows a workflow on a grid background. It consists of three main steps: two "Input dataset" blocks, a "Join" block, and a "Group" block. The "Join" block has a "with" section containing "output (interval)". The "Group" block has a "Select data" section containing "out\_file1 (tabular)". Arrows indicate the flow from the input datasets to the join step, and then to the group step. A small preview window in the bottom right shows a visualization of overlapping features as colored bars.
- Details:** A panel on the right titled "Details" for the selected workflow. It includes:
  - Edit Workflow Attributes:**
    - Name:** count overlapping features
    - Tags:** A section with a tag icon and the text: "Apply tags to make it easy to search for and find items with the same tag."
    - Annotation / Notes:** A section with the text: "Describe or add notes to workflow. Add an annotation or notes to a workflow; annotations are available when a workflow is viewed."

Published Workflow: Count Overlaps Between Feature Sets

# Workflow editor: save your changes

The screenshot displays a workflow editor interface. On the left, a sidebar lists 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 canvas, titled 'Workflow Canvas | count overlapping features', shows a workflow with two 'Input dataset' tools connected to a 'Join' tool. The 'Join' tool has a context menu open over it, with options: Save, Run, Edit Attributes, Auto Re-layout, and Close. The 'Save' option is highlighted. The top bar shows 'Workflow Canvas | count overlapping features' and a settings gear icon. The right sidebar shows 'Details' and 'Edit Workflow Attributes'.

Published Workflow: Feature Overlap Counting

# Workflow Testing

Guided: rerun with same inputs

Workflow → Run

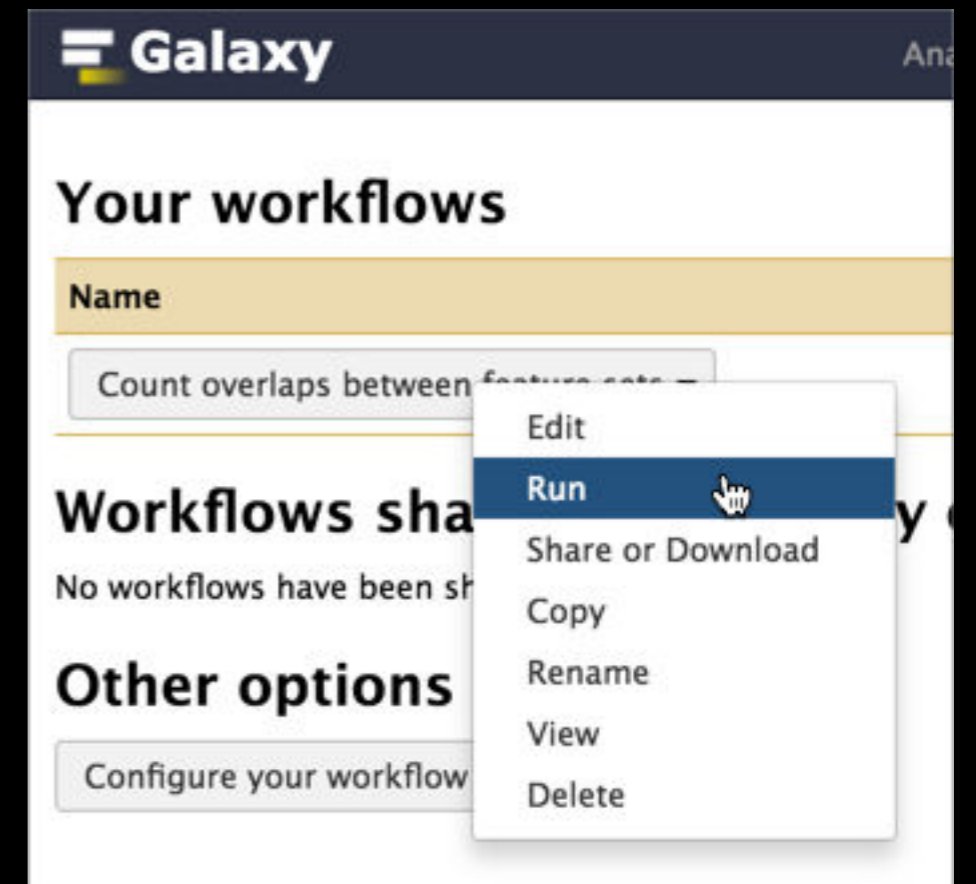
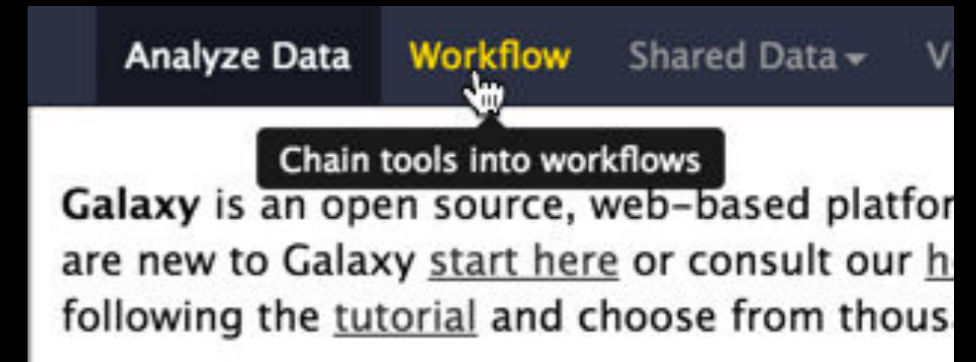
Did that work?

On your own:

Count # of exons overlapping each  
repeat

Did that work? *Why not?*

Edit workflow: doc assumptions



Published Workflow: Count overlaps between feature sets

# 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

5:00 Done



# 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
- 5:00 Done



# 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 2016 UC Davis Using Galaxy for Analysis of RNA-Seq, Exome-Seq, and Variants.

[bit.ly/ucdrnaseq2016](http://bit.ly/ucdrnaseq2016)

# 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.**



# NGS Data Quality: Assessment tools

NGS QC and Manipulation → **FastQC**

Generates summary quality information.

FastQC Read Quality reports (Galaxy Version 0.63) Versions Options

**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 CAAGCAGAAGACGGCATAACGA

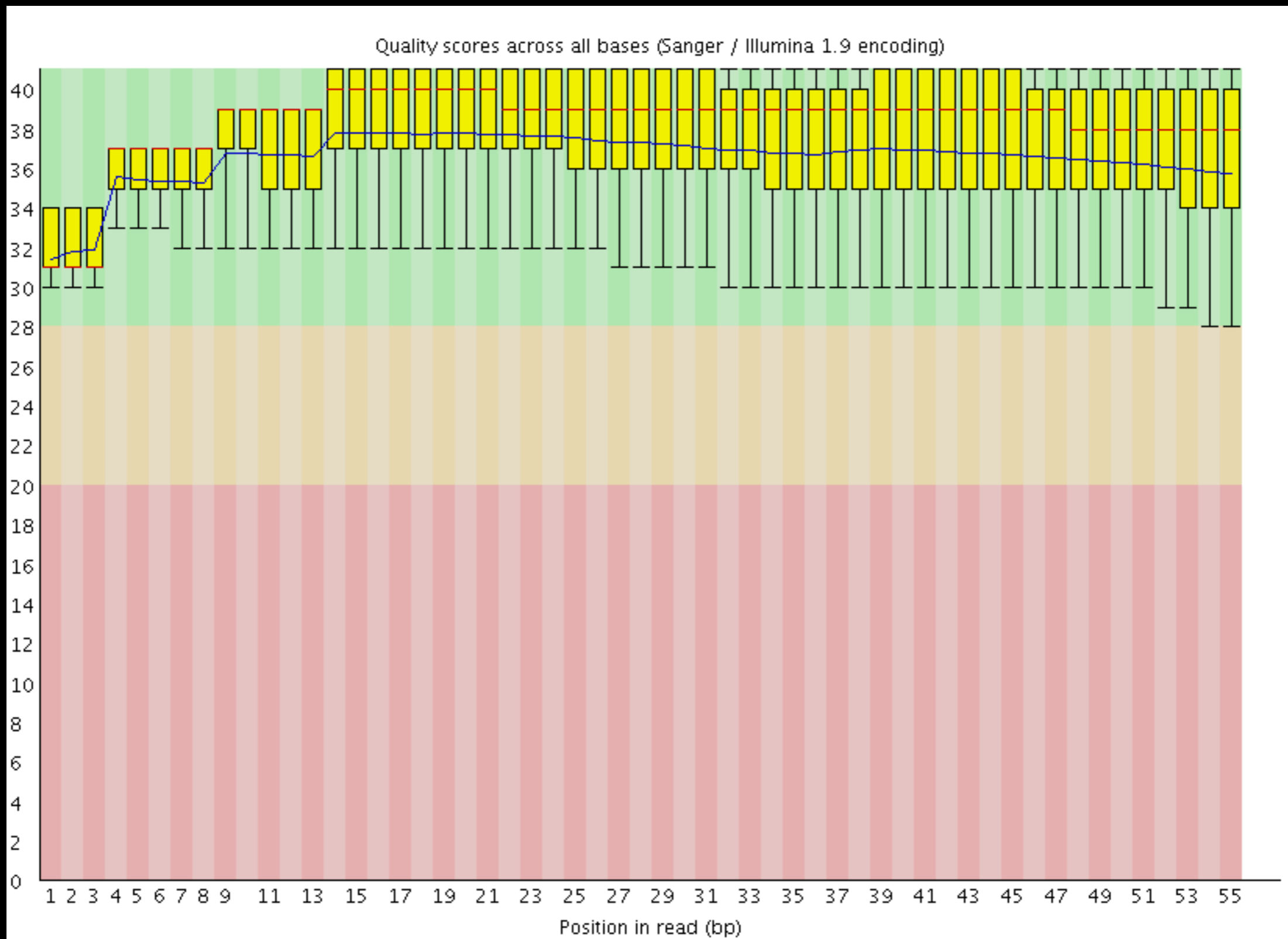
**Submodule and Limit specifying file**

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

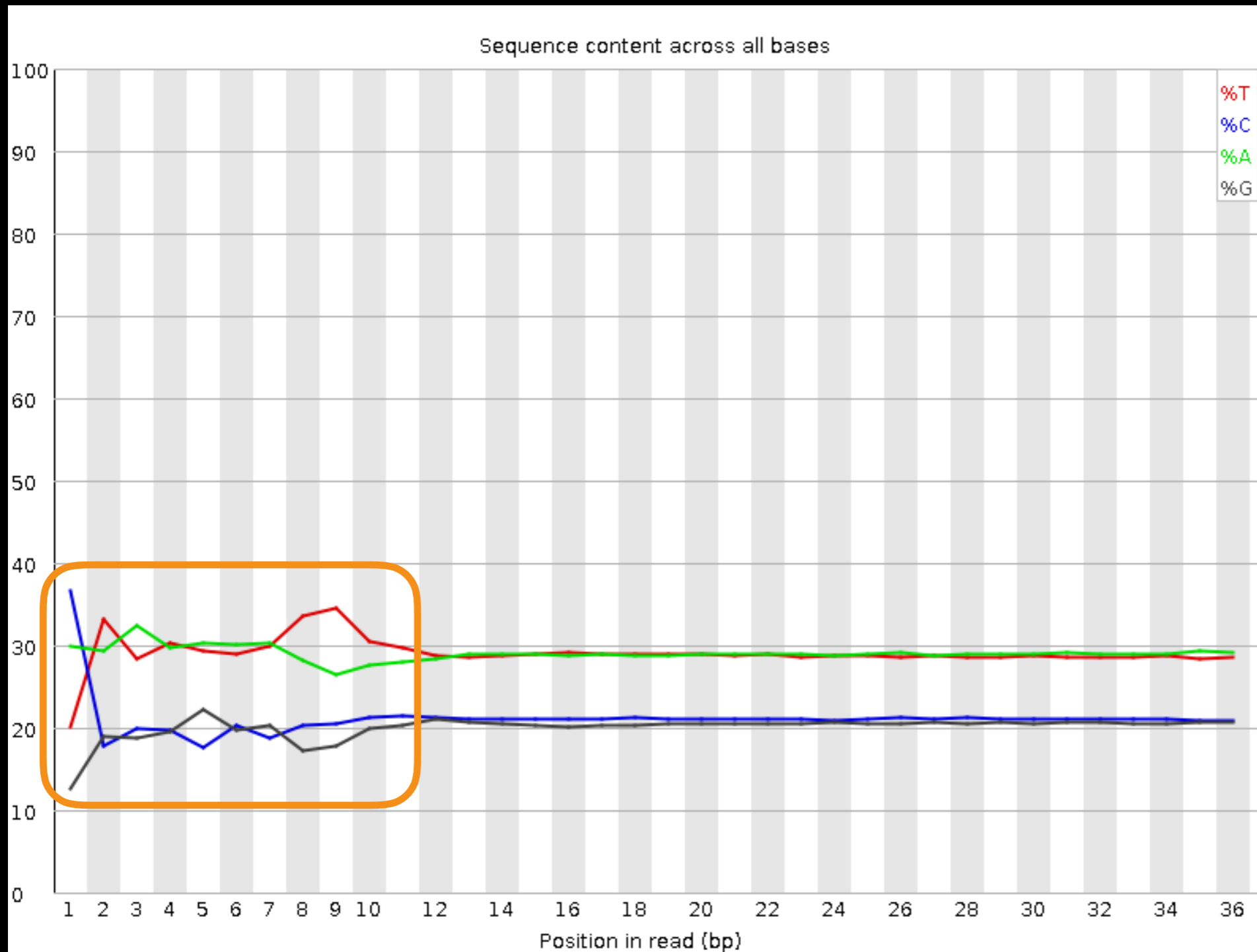
<http://bit.ly/FastQCBoxPlot>

# 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?**  
Yes No

**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?**  
Yes No  
Cut adapter and other illumina-specific sequences from the read

**Trimmomatic Operation**  
1: Trimmomatic Operation trash

**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**.



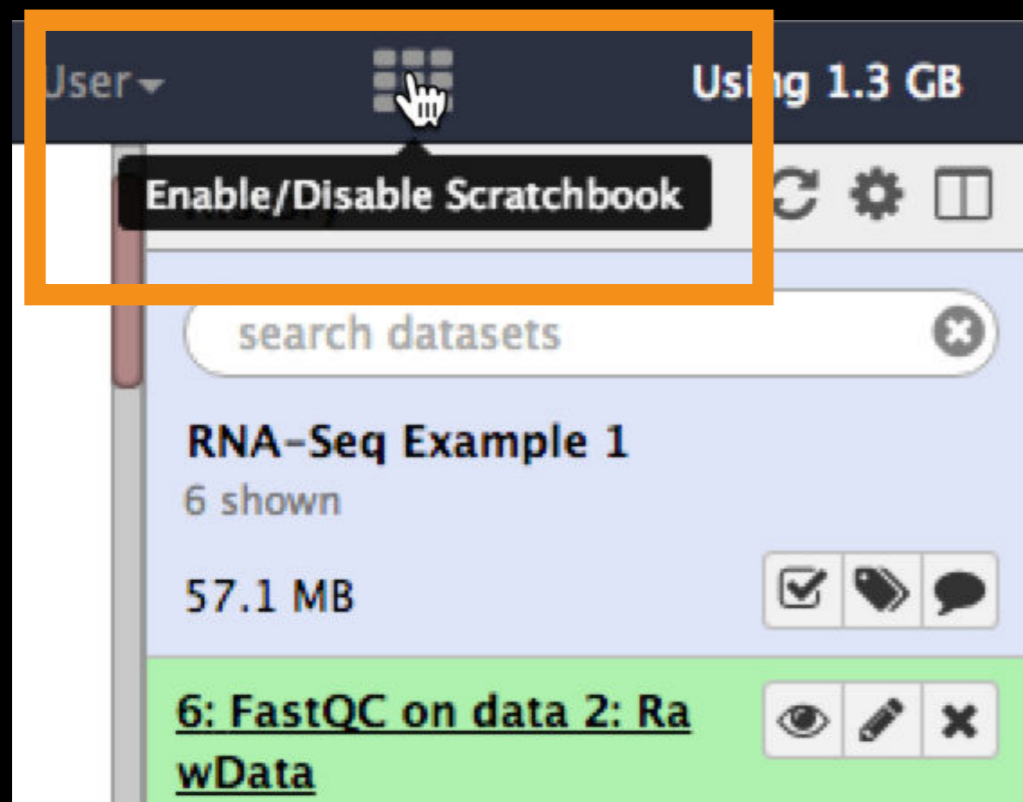
# Run FastQC on post-Trimmatic Datasets

NGS QC and Manipulation → **FastQC**

Now, let's see what changed

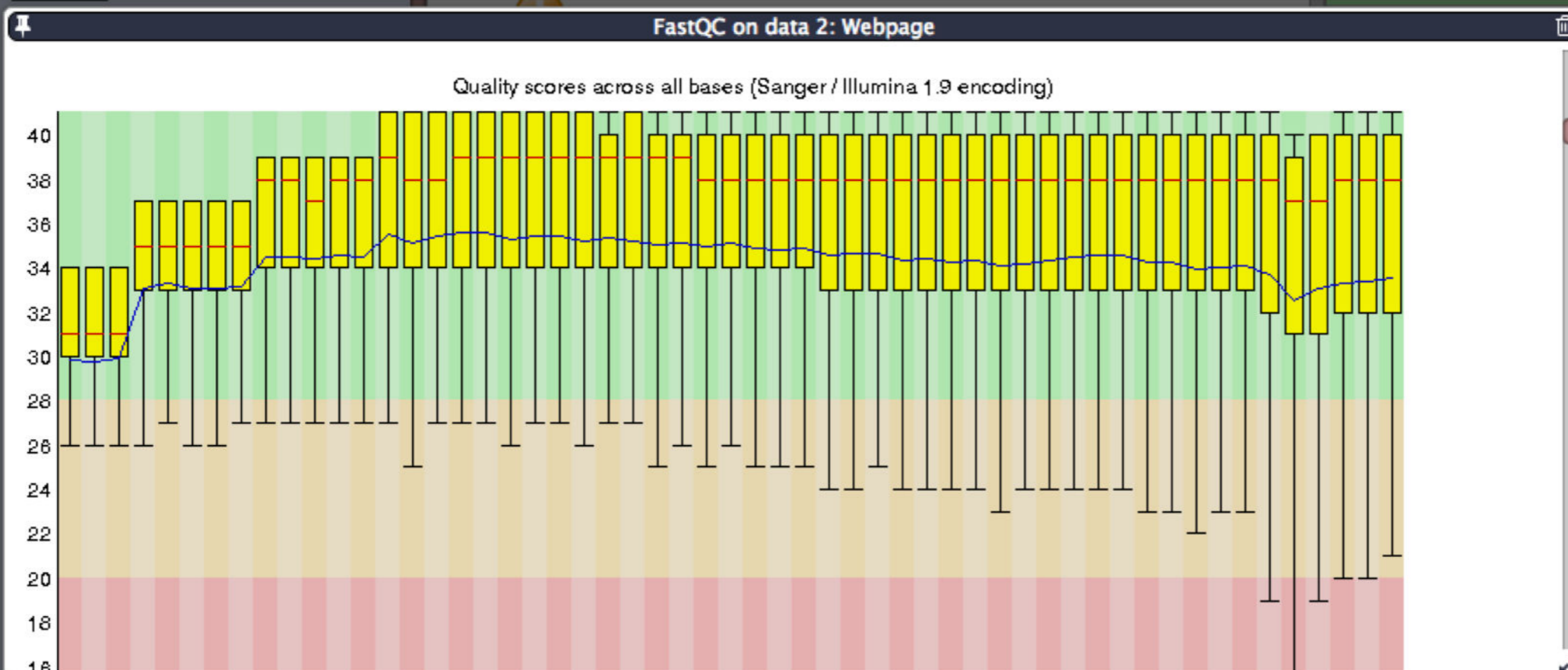
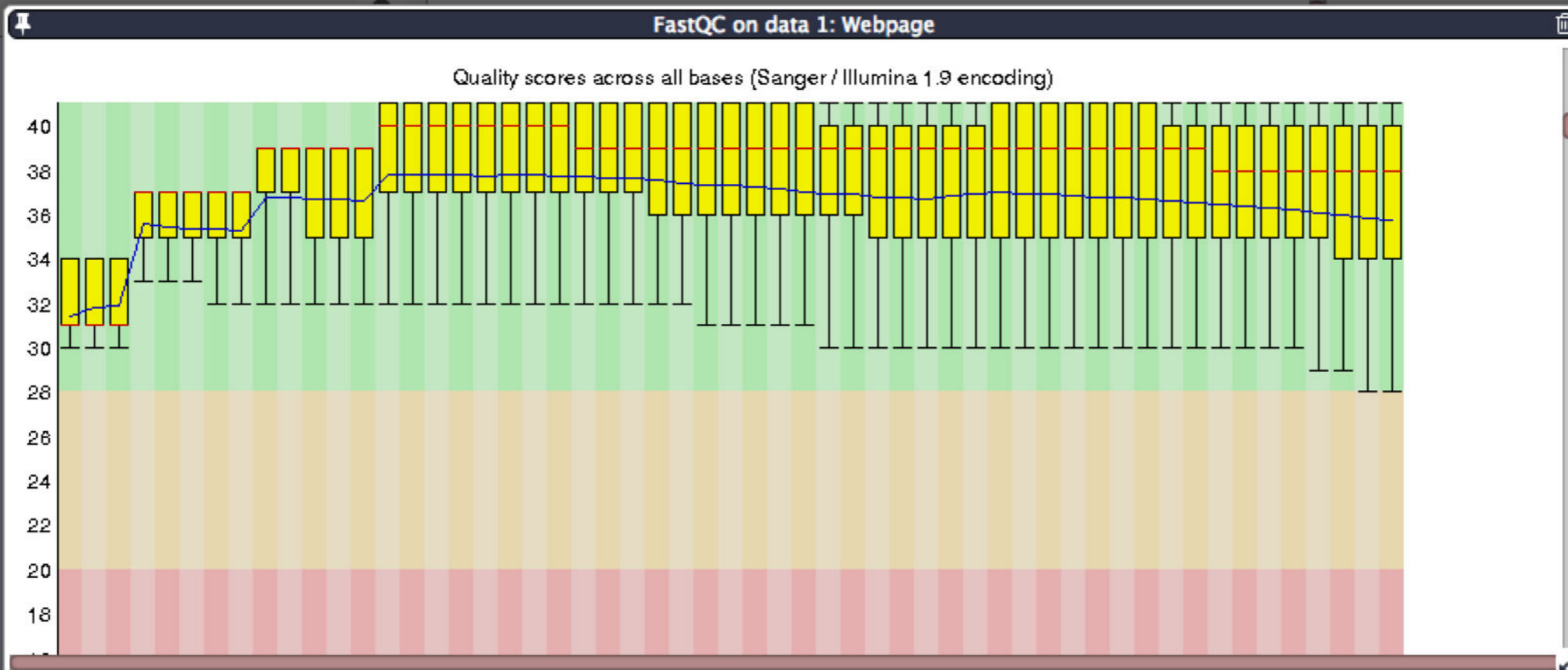
**Shared History: RNA-Seq MeOH\_REPI QC**

# 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

Options

Version 0.32.3)

Paired end data?

Yes

No

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?

Yes

No

Cut adapter and other illumina-specific sequences from the read

Trimmomatic Operation

1: Trimmomatic Operation



Version 0.32.3)

## Paired end data?

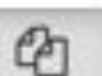
Yes

No

## Input Type

Pair of datasets

## Input FASTQ file (R1/first of pair)



11: R3G\_REP3\_R1.fastq

10: R3G\_REP2\_R2.fastq

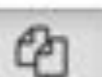
9: R3G\_REP2\_R1.fastq

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**
- 5:00 Done**

# 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**
- 5:00 Done

# 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 chr12.gencode.v25.basic.annotation.gtf

\* RNA-Seq example datasets from the 2016 UC Davis Using Galaxy for Analysis of RNA-Seq, Exome-Seq, and Variants. [bit.ly/ucdrnaseq2016](http://bit.ly/ucdrnaseq2016)

# 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: **Make it quicker?**

**Warning: Here be dragons!**

- **Allow indel search → No**

# 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 →

`chr12.gencode.v25.basic.annotation.gtf`

You can also restrict Tophat to known annotations

- Use Raw Junctions → Yes (tab delimited file)

- Only look for supplied junctions → Yes

# 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)

$\geq 20$

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

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 and Filtered

Select all (OK, maybe just half of them)

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

→ UC-Davis → Reference

Select `chr12.gencode.v25.basic.annotation.gtf`

\* RNA-Seq example datasets from the 2016 UC Davis Using Galaxy for Analysis of RNA-Seq, Exome-Seq, and Variants. [bit.ly/ucdrnaseq2016](http://bit.ly/ucdrnaseq2016)

# Dataset collections!

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

Tools can then take advantage of this knowledge.



# Dataset collections

History ↻ ⚙️ 📄

search datasets ✕

Unnamed history  
7 shown

88.63 MB ☑️ 🗑️ 💬

**7: chr12.gene...  
sic.annotation.gtf** Operations on multiple datasets

**6: R3G REP3 Mapped Filtered** 👁️ ✎ ✕

**5: R3G REP2 Mapped Filtered** 👁️ ✎ ✕

**4: R3G REP1 Mapped Filtered** 👁️ ✎ ✕

**3: MeOH REP3 Mapped Filtered** 👁️ ✎ ✕

**2: MeOH REP2 Mapped Filtered** 👁️ ✎ ✕

**1: MeOH REP1 Mapped Filtered** 👁️ ✎ ✕

History ↻ ⚙️ 📄

search datasets ✕

Unnamed history  
7 shown

88.63 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

**2: MeOH REP2 Mapped Filtered**

**1: MeOH REP1 Mapped Filtered**

# Dataset collections

## Create a collection from a list of datasets

Collections of datasets are permanent, ordered lists of datasets that can be passed to tools and workflows in ... [More help](#)

[Start over](#)

[MeOH\\_REP3\\_Mapped\\_Filtered](#)

Discard

[MeOH\\_REP2\\_Mapped\\_Filtered](#)

Discard

[MeOH\\_REP1\\_Mapped\\_Filtered](#)

Discard

Name:

Cancel

Create list

Thank you for using Galaxy.

# Dataset collections




History   

search datasets 

Unnamed history  
9 shown







88.63 MB   

- 9: R3G   
a list of datasets
- 8: MeOH   
a list of datasets
- 7: chr12.gencode.v25.bas  
sic.annotation.gtf   
- 6: R3G REP3 Mapped Filte  
red   
- 5: R3G REP2 Mapped Filte  
red   
- 4: R3G REP1 Mapped Filte  
red   
- 3: MeOH REP3 Mapped F  
iltered   
- 2: MeOH REP2 Mapped F  
iltered   
- 1: MeOH REP1 Mapped F  
iltered   

History   

< [Back to Unnamed history](#)

MeOH  
a list of datasets

- MeOH REP3 Mapped Filtere  
d  
- MeOH REP2 Mapped Filtere  
d  
- MeOH REP1 Mapped Filtere  
d  

# 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



(Galaxy Version 2.2.1.2)

### Transcripts

  7: chr12.gencode.v25.basic.annotation.gtf

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


  8: MeOH

#### 2: Condition

##### Name

R3G

##### Replicates

  9: R3G

# 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
- 5:00 Done



# The Galaxy Team



Enis Afgan



Dannon Baker



Dan Blankenberg



Dave Bouvier



Marten Cech



John Chilton



Dave Clements



Nate Coraor



Jeremy Goecks



Sam Guerler



Mo Heydarian



Jen Jackson



Vahid Jalili



Delphine Lariviere



Ross Lazarus



Anton Nekrutenko



Nick Stoler



James Taylor

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

# Acknowledgements

You  
Karen Eilbeck

Department of Biomedical Informatics  
University of Utah

NIH  
AWS  
Johns Hopkins University  
Penn State University

[bit.ly/btigxy\\_feedback](https://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
- 5:00 Done**





26 - 30 June France

# GCC 2017 Montpellier



Le Corum  
Conference centre

[gcc2017.sciencesconf.org](http://gcc2017.sciencesconf.org)

# 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

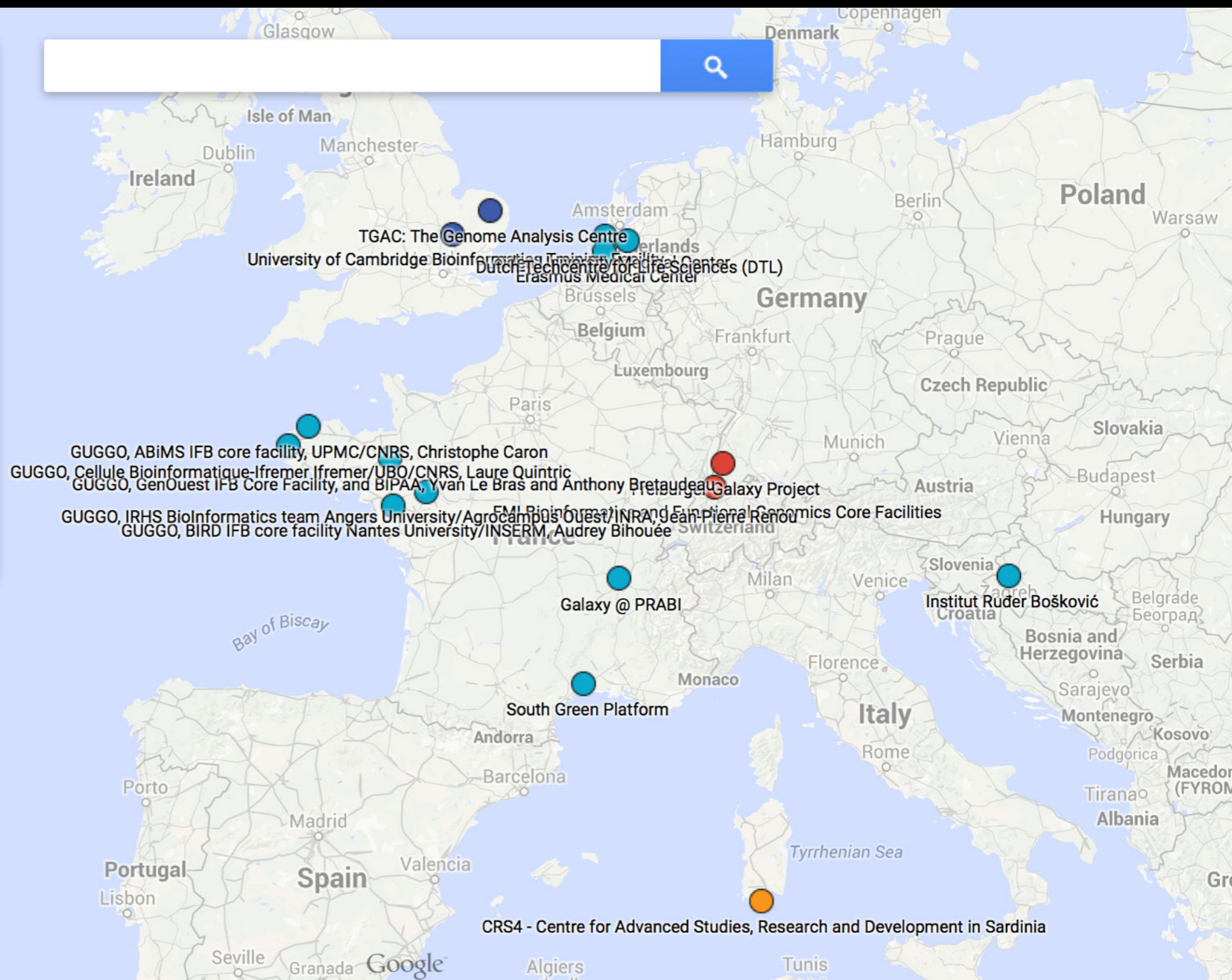
The Galaxy Training Network  
(<https://wiki.galaxyproject.org/Teach/GTN>)



Made with Google My Maps

### Trainers

- Global
- Regional
- Local
- Continental
- Institution



## Galaxy Training Network

[bit.ly/gxygtn](https://bit.ly/gxygtn)





# Galaxy Community Resources: Mailing Lists

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

## Galaxy-Dev

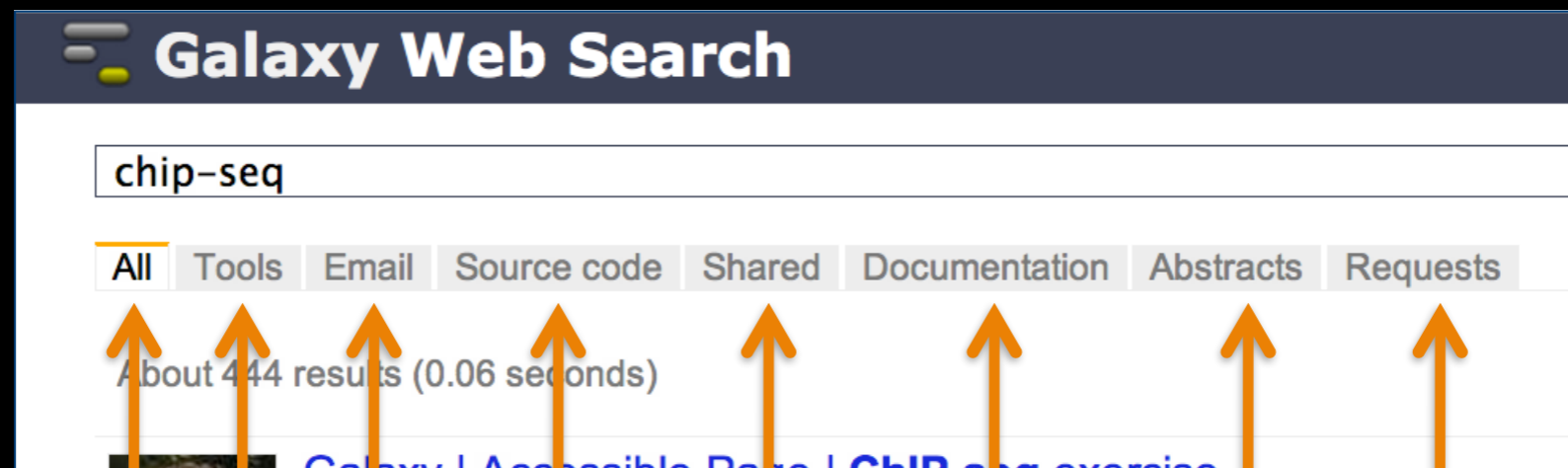
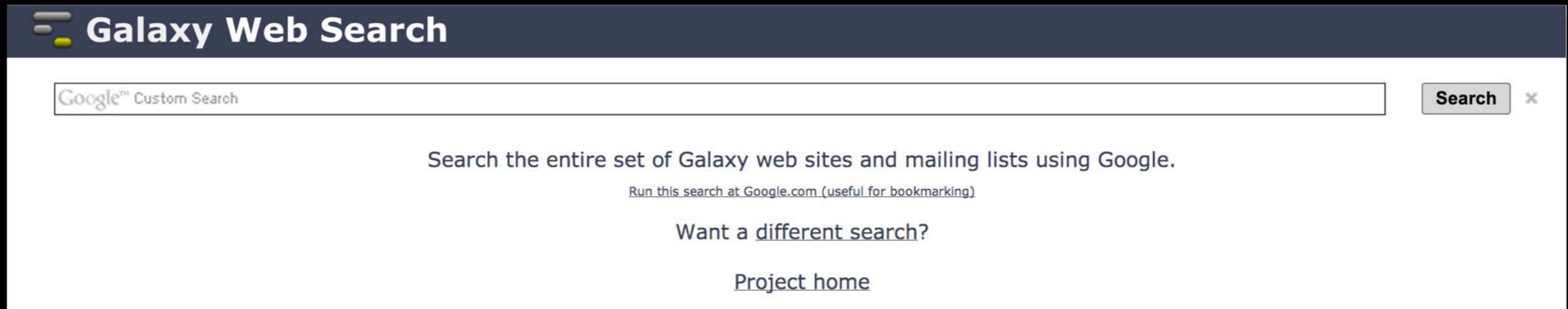
Questions about developing for and deploying Galaxy  
High volume (2336 posts in 2015, 1000+ members)

## 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](http://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

## 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@glaxyproject.org](mailto:outreach@glaxyproject.org).

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

## Upcoming Events



Date	Topic/Event	Venue/Location
December 12	<a href="#">Introduction to Galaxy Workshop</a>	Virginia State University, Petersburg, Virginia
December 16-19	<a href="#">RNA-Seq and ChIP-Seq Analysis with Galaxy</a>	UC Davis, California, United States
<b>2015</b>		
January 10-14	<a href="#">Galaxy for SNP and Variant Data Analysis</a>	Plant and Animal Genome XXIII (PAG2014), States
January 19-20	<a href="#">NGS pipelines with Galaxy</a>	e-Infrastructures for Massively Parallel Sequencing, Sweden
February 9-13	<a href="#">Analyse bioinformatique de séquences sous Galaxy</a>	Montpellier, France
February 16-18	<a href="#">Accessible and Reproducible Large-Scale Analysis with Galaxy</a>	Genome and Transcriptome Analysis, Pacific Conference, San Francisco, California
	<a href="#">Large-Scale NGS data Analysis on Amazon Web Services Using Globus Genomic iReport: An Integrative "omics"</a>	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 controls

# Galaxy Resources & Community: Videos

**vimeo** Me Videos Create Watch Tools Upload Search

## 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
- usegalaxy.org**  
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



CiteULike MyCiteULike Group: Galaxy Search Logged in as galaxyproject Log Out

## Group: Galaxy - library 3726 articles

You are an administrative member of this group.  
Invite [other CiteULike users](#) to join, or invite [people who don't use CiteULike yet](#).

Search Unwatch Copy Export Sort Hide Details

- Y-box protein 1 is required to sort microRNAs into exosomes in cells and in a**  
*eLife*, Vol. 5 (25 August 2016), [doi:10.7554/elife.19276](https://doi.org/10.7554/elife.19276)  
by [Matthew J. Shurtleff](#), [Morayma M. Temoche-Diaz](#), [Kate V. Karfilis](#), [Sayaka Ri](#), [Randy Sch](#)  
posted to [methods](#) [usemain](#) by [galaxyproject](#) to the group [Galaxy](#) keyed Shurtleff2016Ybox  
■ Copy ■ My Copy
- Validation and characterization of thirteen microsatellite markers for queen con**  
*PeerJ Preprints*, Vol. 4 (October 2016), [doi:10.7287/peerj.preprints.2559v1](https://doi.org/10.7287/peerj.preprints.2559v1)  
by [Nathan K. Truelove](#), [Loong Fai Ho](#), [Richard F. Preziosi](#), [Stephen J. Box](#)  
posted to [methods](#) [uselocal](#) by [galaxyproject](#) to the group [Galaxy](#) keyed 10.7287/peerj.prep  
03:36:18 ★★/  
■ Abstract ■ Copy
- Transcriptomic analysis reveals how a lack of potassium ions increases Sulfo**  
**sensitivity to pH changes**  
*Microbiology*, Vol. 162, No. 8. (01 August 2016), pp. 1422-1434, [doi:10.1099/mic.0.000314](https://doi.org/10.1099/mic.0.000314)  
by [Antoine Buetti-Dinh](#), [Ran Friedman](#), [Olga Dethlefsen](#), [Mark Dopson](#)  
posted to [methods](#) by [galaxyproject](#) to the group [Galaxy](#) keyed BuettiDinh2016Transcripto  
■ Copy ■ My Copy

### Group Tags

All tags in the group Galaxy

Filter:

[\[Display as Cloud\]](#)

<a href="#">methods</a>	1864
<a href="#">workbench</a>	1030
<a href="#">usemain</a>	397
<a href="#">usepublic</a>	373
<a href="#">tools</a>	258
<a href="#">isgalaxy</a>	194
<a href="#">uselocal</a>	184
<a href="#">refpublic</a>	164
<a href="#">cloud</a>	144
<a href="#">other</a>	121
<a href="#">shared</a>	105
<a href="#">reproducibility</a>	97
<a href="#">unknown</a>	72
<a href="#">howto</a>	65
<a href="#">project</a>	54
<a href="#">visualization</a>	27
<a href="#">usecloud</a>	7

Now almost 4000 papers

<http://bit.ly/gxycul>

