Introduction to Galaxy

Virginia State University
December 12, 2014

Dave Clements
Galaxy Project
Johns Hopkins University

Morning Agenda

10:00	Welcome: Introduction and Logistics
10:15	Basic analysis with Galaxy
11:40	Galaxy Project Resources
12:00	Lunch (catered)
1:00	Advanced Usage: RNA-Seq Analysis
3:00	Done

Goals

Provide a basic 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.

Not Goals

This workshop will *not* cover

- details of how tools are implemented, or
- new algorithm designs, or
- which assembler or mapper or peak caller or ... is best for you.

While this workshop does cover Galaxy you won't become a Galaxy expert in the next two hours.

What is Galaxy?

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

A free (for everyone) web server

Open source software

These options result in several ways to use Galaxy

http://galaxyproject.org

Galaxy is available ...

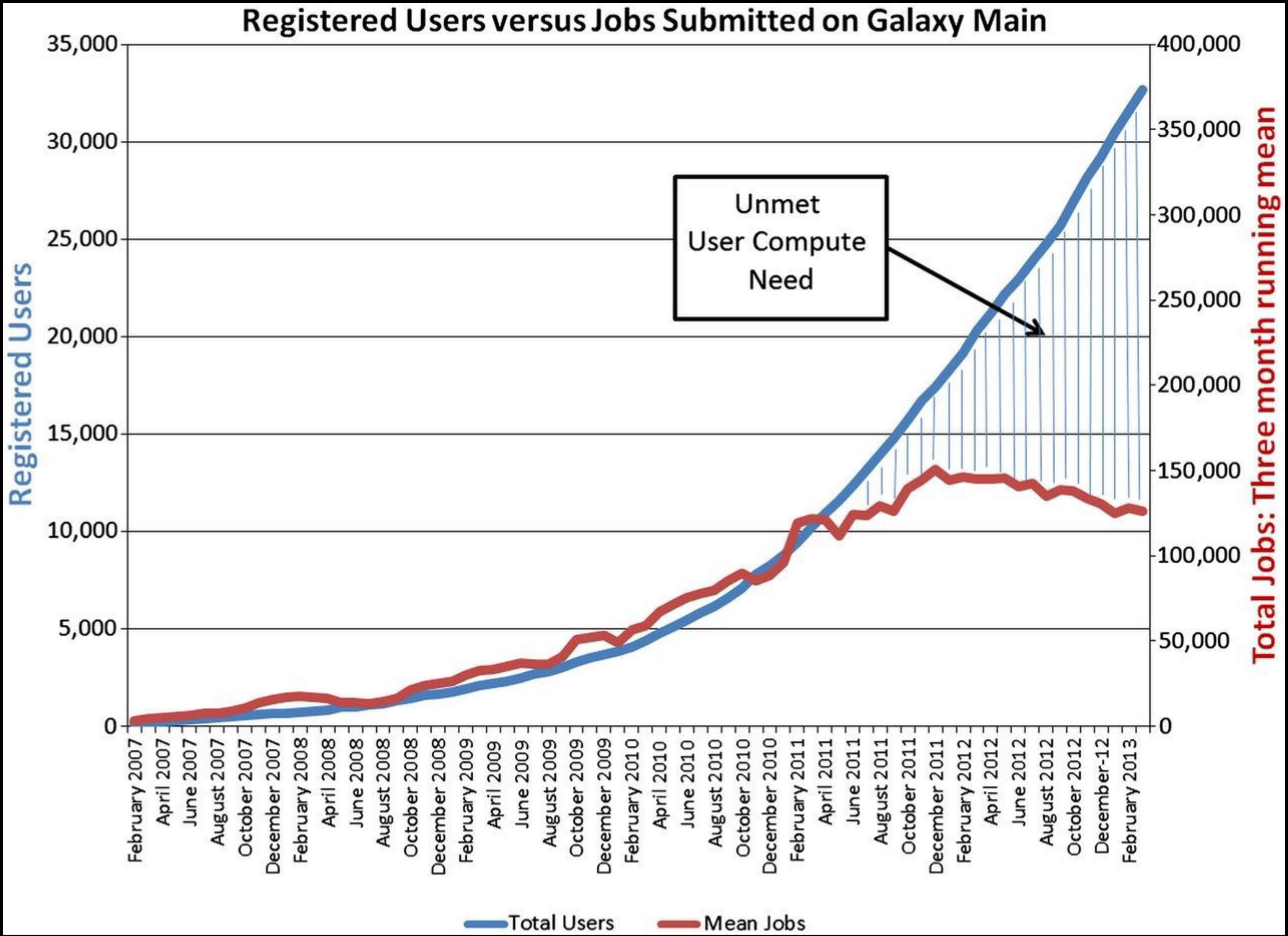
As a free (for everyone) web server integrating a wealth of tools, compute resources, petabytes of reference data and permanent storage

http://usegalaxy.org

TACC

iPlant Collaborative™

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



Leveraging the national cyberinfrastructure for biomedical research

LeDuc, *et al. J Am Med Inform Assoc doi:10.1136/amiajnl-2013-002059*

Galaxy is available ...

- As a free (for everyone) web service

  http://usegalaxy.org

- As open source software

  http://getgalaxy.org

  It is installed in locations around the world

Galaxy is available ...

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

We are using the cloud today.

Galaxy is available: With Commercial Support

A ready-to-use appliance

(BioTeam)

Cloud-based solutions

(ABgenomica, AIS, GenomeCloud)

Consulting & Customization

(Arctix, BioTeam, Deena Bioinformatics)

Galaxy Project: Further reading & Resources

http://galaxyproject.org

http://usegalaxy.org

http://getgalaxy.org

http://wiki.galaxyproject.org/Cloud

http://bit.ly/gxychoices

Agenda

10:00	Welcome: Introduction and Logistics
10:15	Basic analysis with Galaxy
11:40	Galaxy Project Resources
12:00	Lunch (catered)
1:00	Advanced Usage: RNA-Seq Analysis
3:00	Done

Basic Analysis

Which exons have most overlapping Repeats?

Use Human, HG19, Chromosome 22

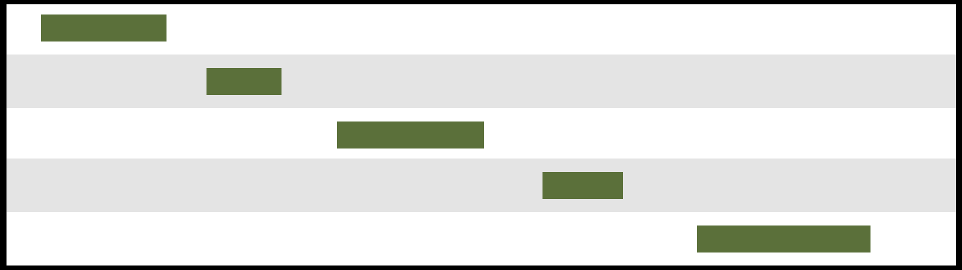
cloud1.galaxyproject.org
cloud2.galaxyproject.org

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

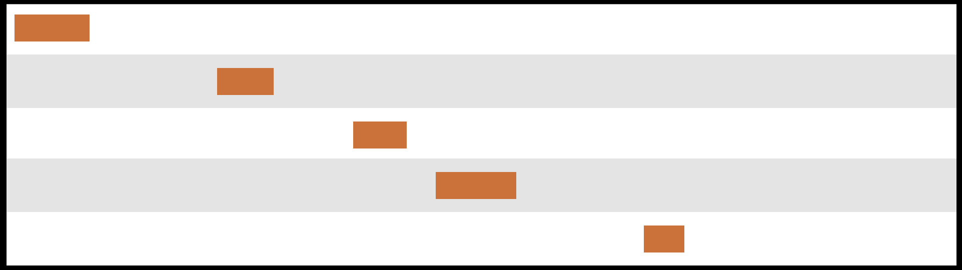
Exons & Repeats: A General Plan

- 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)

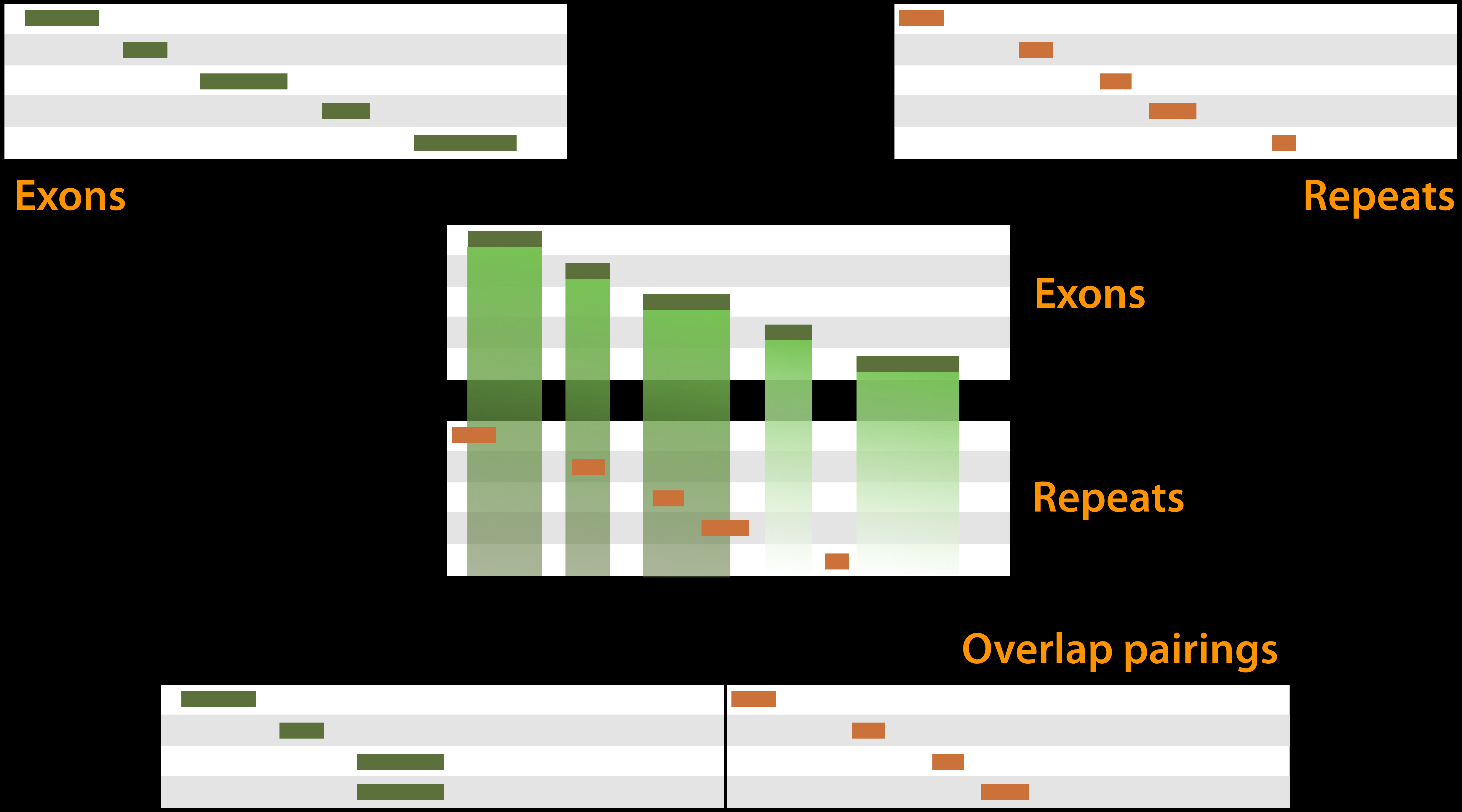


Exons



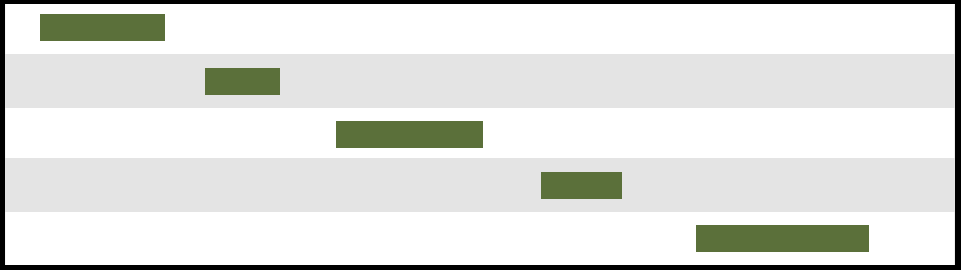
Repeats

(Identify which exons have Repeats)

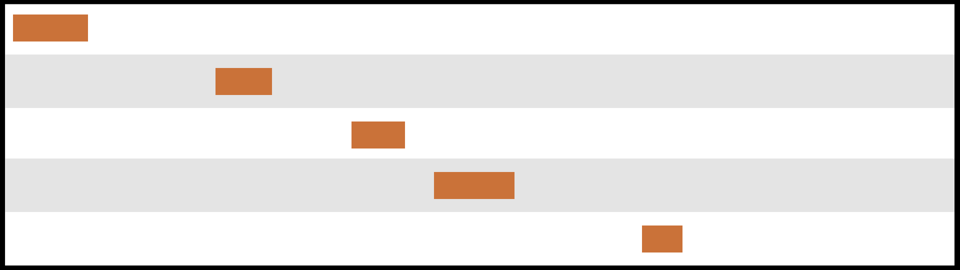


Operate on Genomic Intervals → **Join**

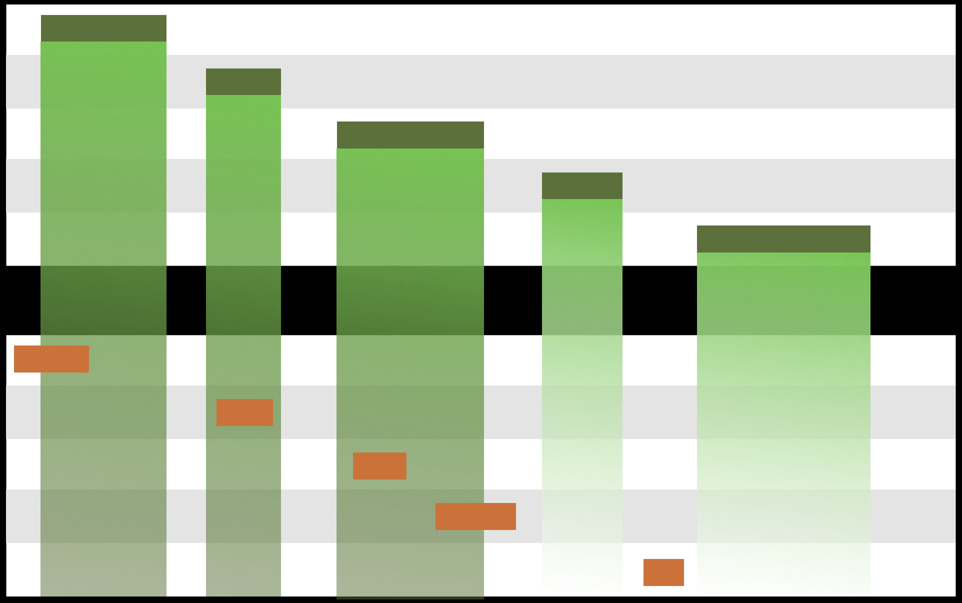
(Identify which exons have Repeats)



Exons



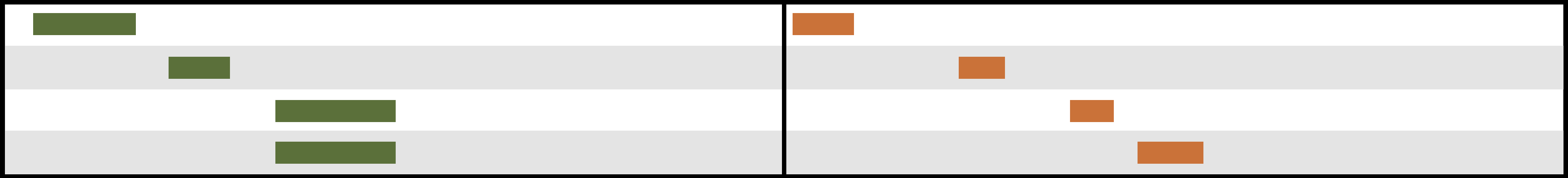
Repeats

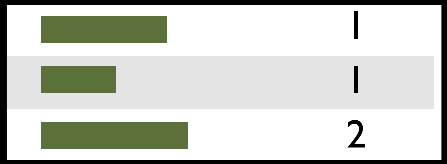


Exons

Repeats

Overlap pairings

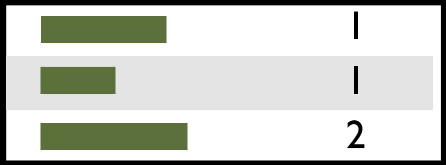




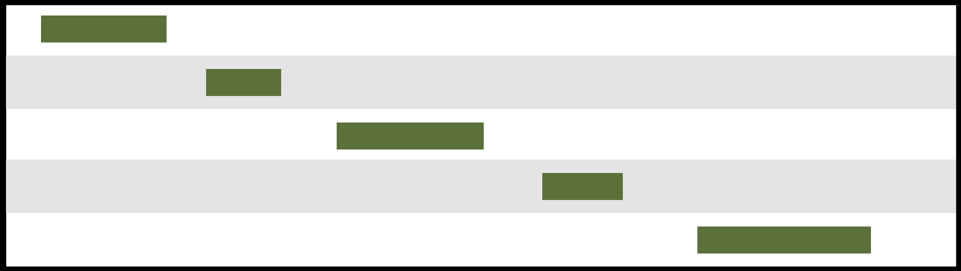
Exon overlap counts

Join, Subtract, and Group → Group

(Count Repeats per exon)



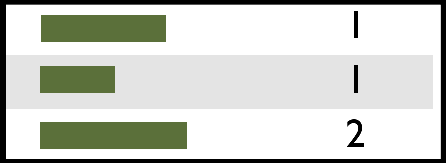
Exon overlap counts



Exons

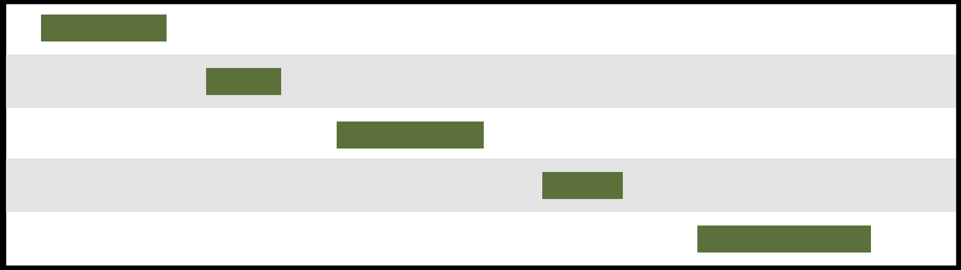
We’ve answered our question, but we can do better.

Incorporate the overlap count with rest of Exon information



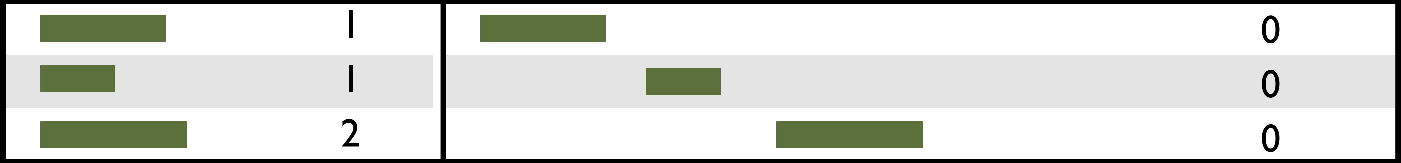
🟩	1
🟩	1
🟩	2

Exon overlap counts



🟩
🟩
🟩
🟩
🟩

Exons

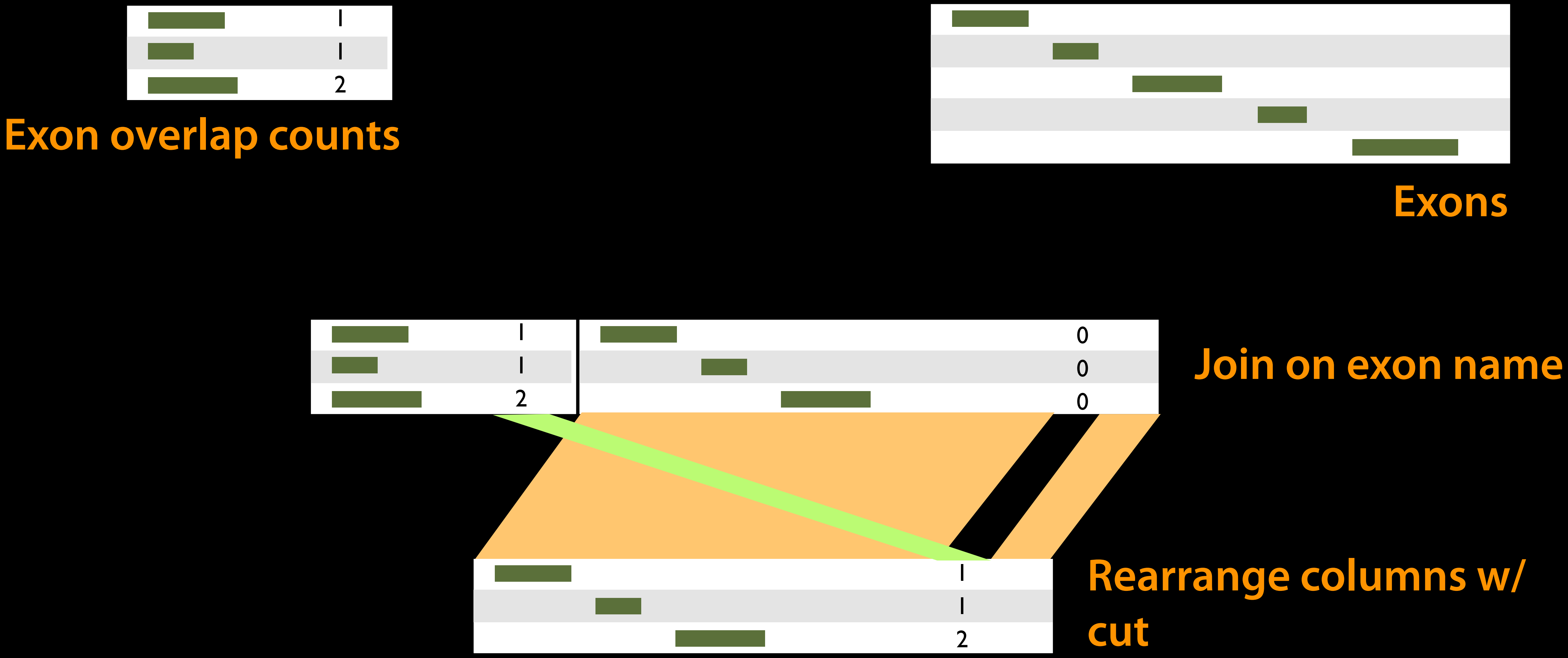


🟩	1	🟩	0
🟩	1	🟩	0
🟩	2	🟩	0

Join on exon name

Join, Subtract, and Group → Join

(Incorporate the overlap count with rest of Exon information)



Exon overlap counts

Exons

Join on exon name

Rearrange columns w/ cut

Text Manipulation → Cut

(Incorporate the overlap count with rest of Exon information)

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
- But, ...
  - there is nothing inherent in the analysis about humans, exons or repeats
  - It is a series of steps that sets the score of one set of features to the number of overlaps from another set of features.

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

Run / test it

Guided: rerun with same inputs

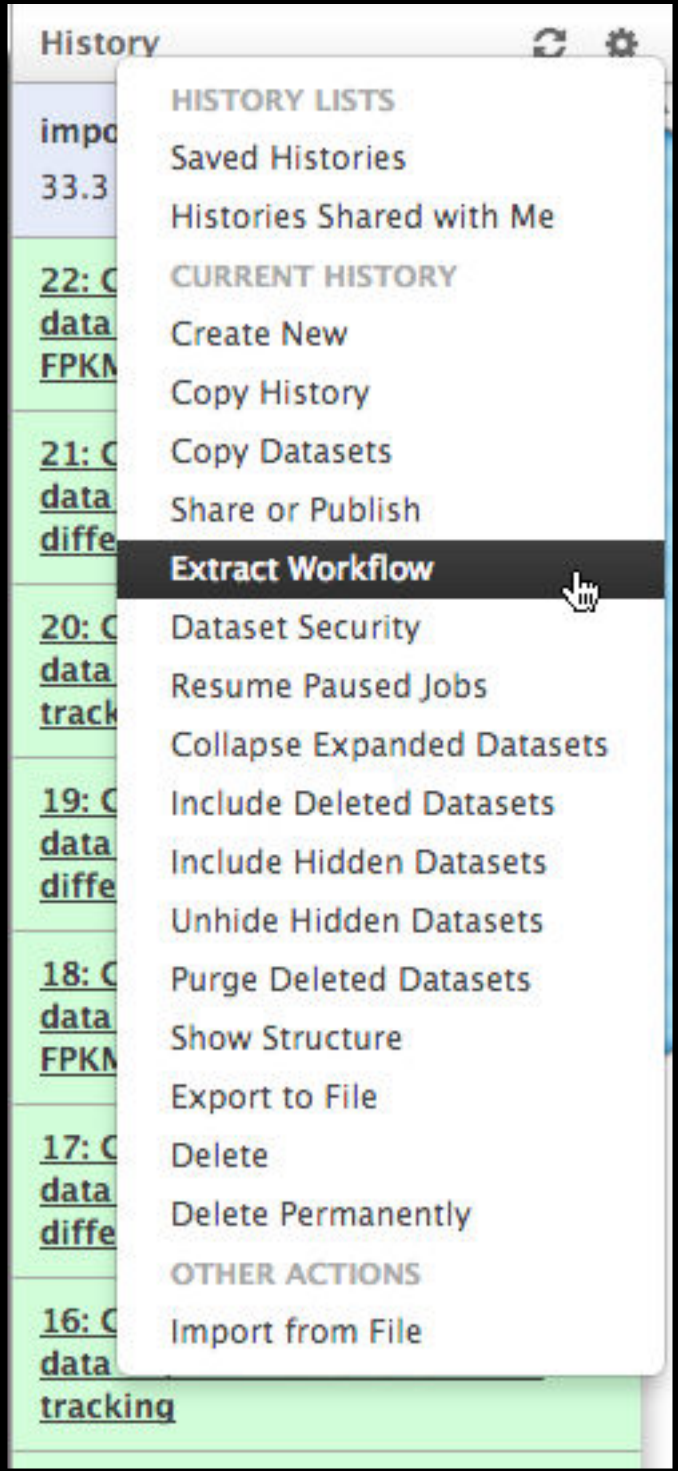
Did that work?

On your own:

Count # of exons in each Repeat

Did that work? *Why not?*

Edit workflow: doc assumptions



More Galaxy Terminology

Share:

Make something available to someone else

Publish:

Make something available to everyone

Galaxy Page:

Analysis documentation within Galaxy; easy to embed any Galaxy object

Let’s all share...

Sharing & Publishing enables Reproducibility

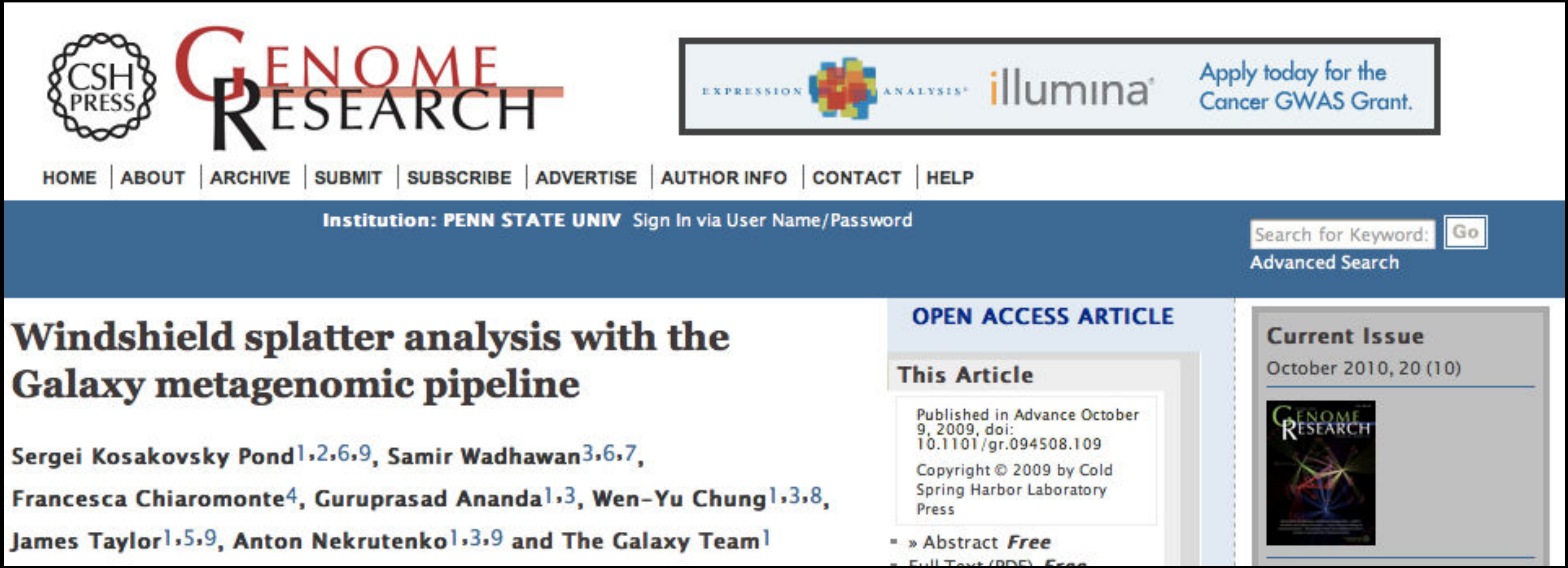
Reproducibility: Everybody talks about it, but ...

Galaxy aims to push the goal of reproducibility from the bench to the bioinformatics realm

All analysis in Galaxy is recorded without any extra effort from the user.

Histories, workflows, visualizations and ***pages*** can be shared with others or published to the world.

Sharing & Publishing enables Reproducibility

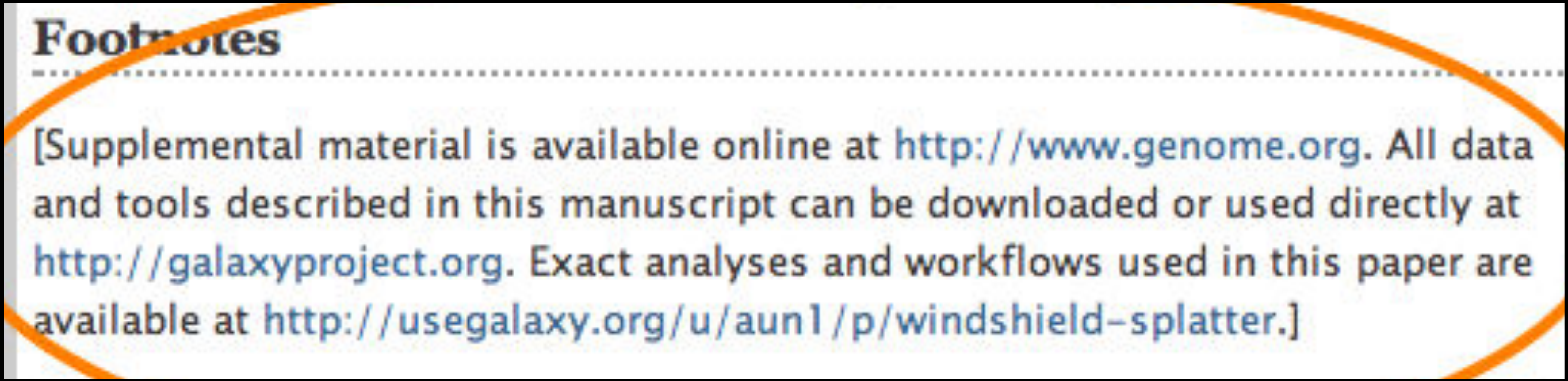

CSH PRESS
GENOME RESEARCH
EXPRESSION ANALYSIS illumina Apply today for the Cancer GWAS Grant.
HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CONTACT | HELP
Institution: PENN STATE UNIV Sign In via User Name/Password
Search for Keyword:
Go
Advanced Search
OPEN ACCESS ARTICLE
This Article
» Abstract Free
» Full Text (PDF) Free
Current Issue
October 2010, 20 (10)

Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond[1,2,6,9], Samir Wadhawan[3,6,7], Francesca Chiaromonte[4], Guruprasad Ananda[1,3], Wen-Yu Chung[1,3,8], James Taylor[1,5,9], Anton Nekrutenko[1,3,9] and The Galaxy Team[1]

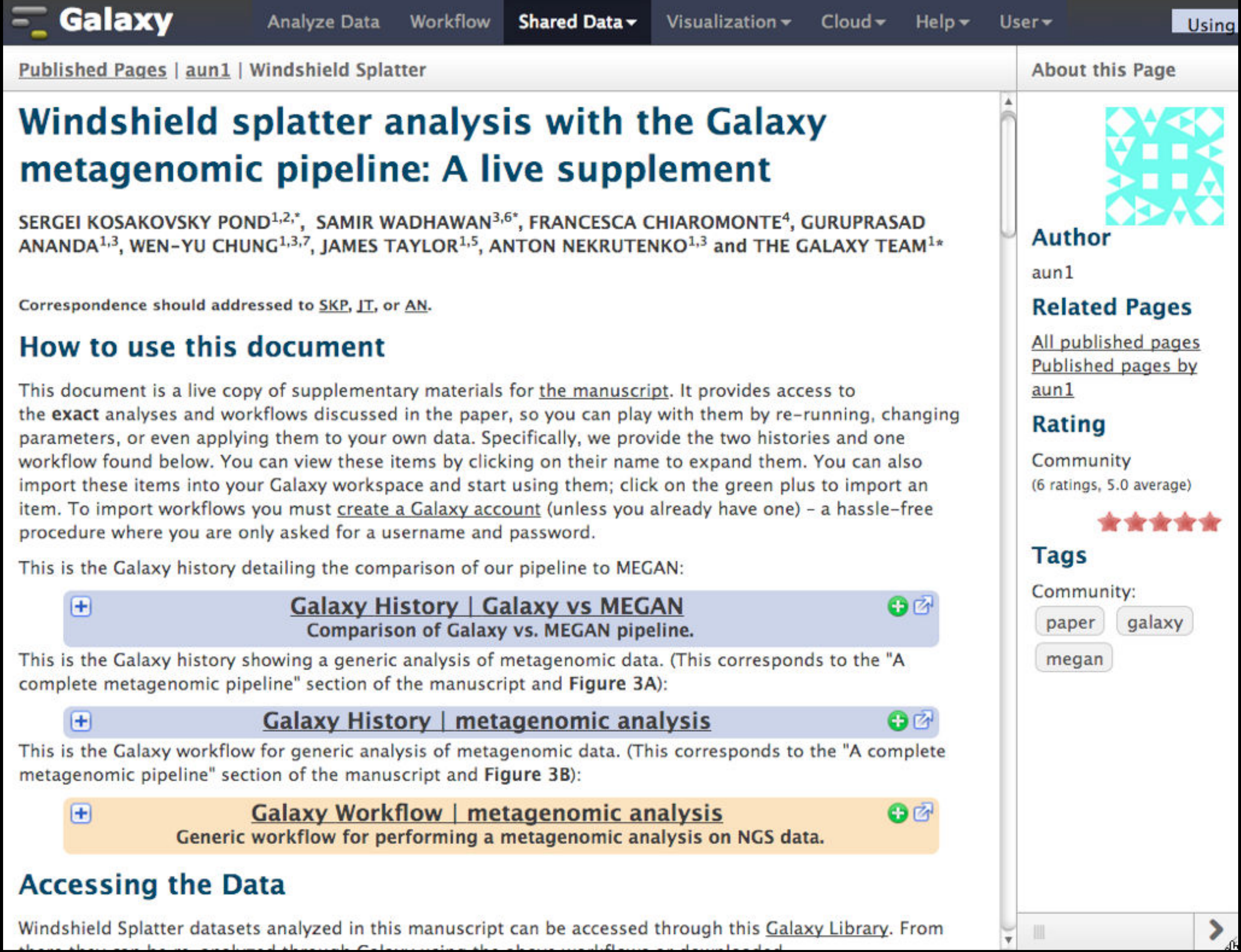
Published in Advance October 9, 2009, doi: 10.1101/gr.094508.109

Copyright © 2009 by Cold Spring Harbor Laboratory Press



Footnotes

[Supplemental material is available online at http://www.genome.org. All data and tools described in this manuscript can be downloaded or used directly at http://galaxyproject.org. Exact analyses and workflows used in this paper are available at http://usegalaxy.org/u/aun1/p/windshield-splatter.]



Published Pages | aun1 | Windshield Splatter

Windshield splatter analysis with the Galaxy metagenomic pipeline: A live supplement

SERGEI KOSAKOVSKY POND[1,2,*], SAMIR WADHAWAN[3,6*], FRANCESCA CHIAROMONTE[4], GURUPRASAD ANANDA[1,3], WEN-YU CHUNG[1,3,7], JAMES TAYLOR[1,5], ANTON NEKRUTENKO[1,3] and THE GALAXY TEAM[1*]

Correspondence should addressed to SKP, JT, or AN.

How to use this document

This document is a live copy of supplementary materials for the manuscript. It provides access to the **exact** analyses and workflows discussed in the paper, so you can play with them by re-running, changing parameters, or even applying them to your own data. Specifically, we provide the two histories and one workflow found below. You can view these items by clicking on their name to expand them. You can also import these items into your Galaxy workspace and start using them; click on the green plus to import an item. To import workflows you must create a Galaxy account (unless you already have one) - a hassle-free procedure where you are only asked for a username and password.

This is the Galaxy history detailing the comparison of our pipeline to MEGAN:

Galaxy History | Galaxy vs MEGAN
Comparison of Galaxy vs. MEGAN pipeline.

This is the Galaxy history showing a generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3A**):

Galaxy History | metagenomic analysis

This is the Galaxy workflow for generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3B**):

Galaxy Workflow | metagenomic analysis
Generic workflow for performing a metagenomic analysis on NGS data.

Accessing the Data

Windshield Splatter datasets analyzed in this manuscript can be accessed through this Galaxy Library. From

About this Page

Author

aun1

Related Pages

All published pages
Published pages by aun1

Rating

Community
(6 ratings, 5.0 average)
★★★★★

Tags

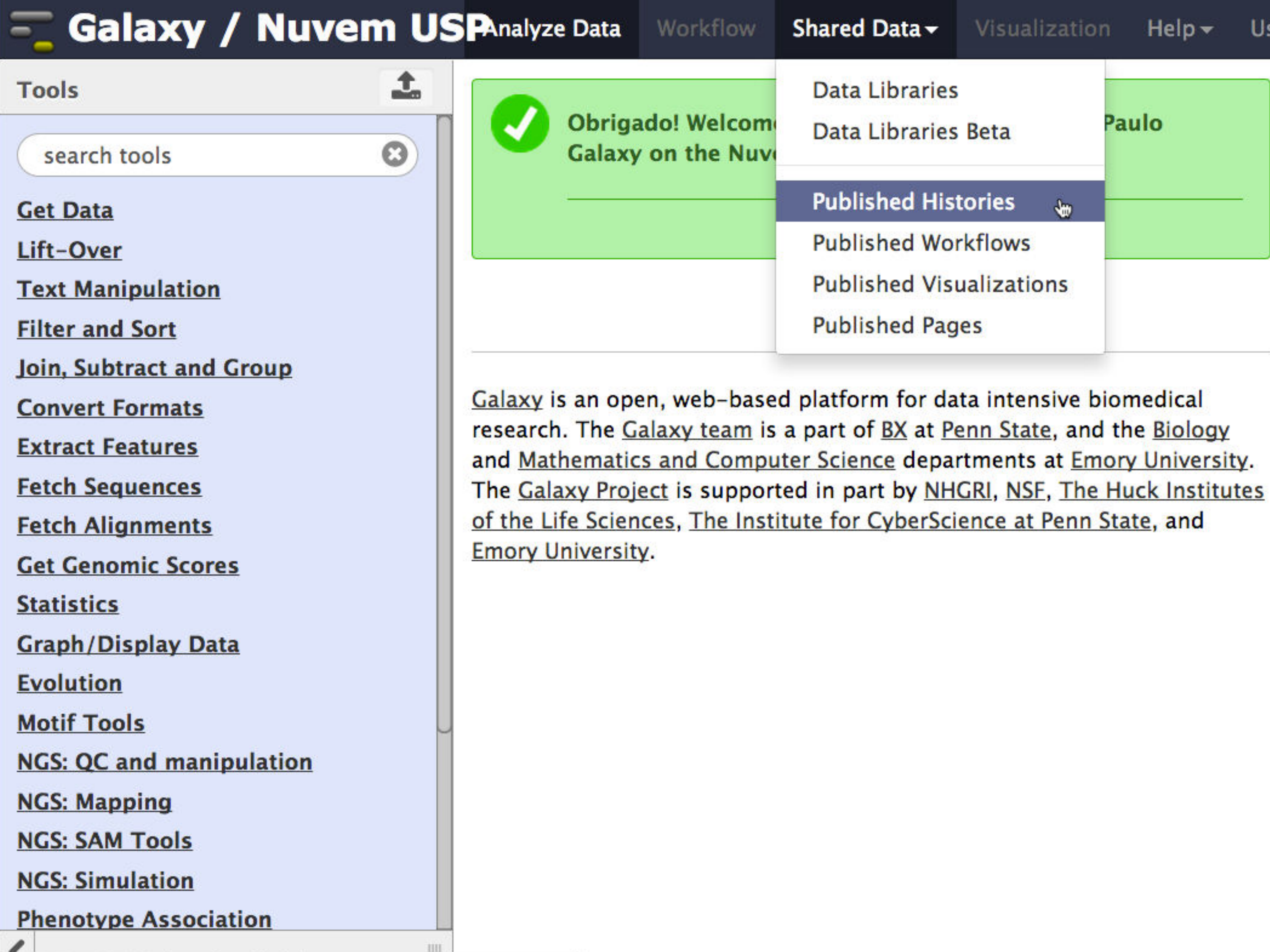
Community:
paper
galaxy
megan

http://usegalaxy.org/u/aun1/p/windshield-splatter

Exons & Repeats: Exercise

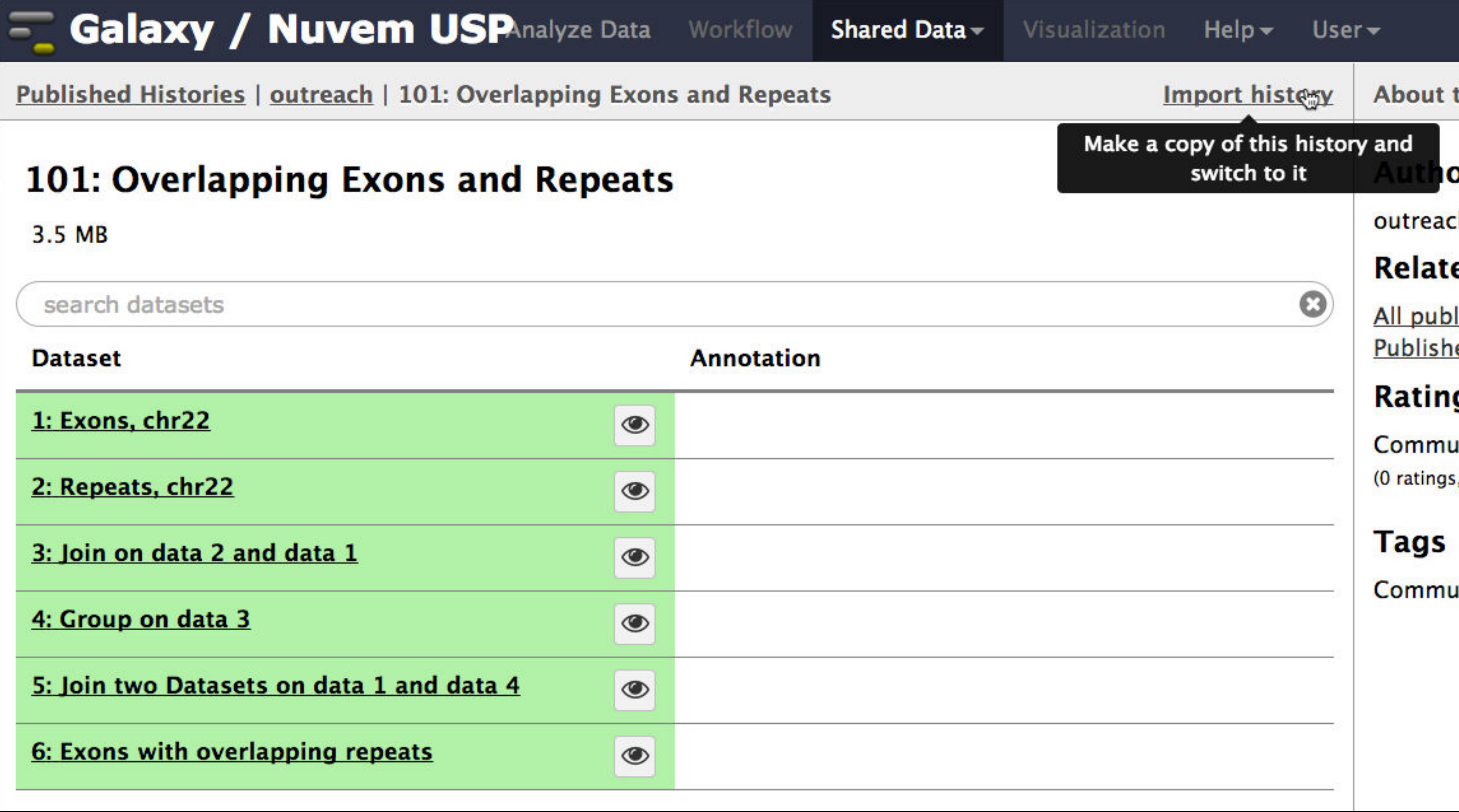
Include exons with no overlaps in final output.
Set the score for these to 0.

Everything you need will be in the toolboxes we used in the Exon-Repeats exercise.



Obrigado! Welcom Paulo
Galaxy on the Nuv

Galaxy is an open, web-based platform for data intensive biomedical research. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.



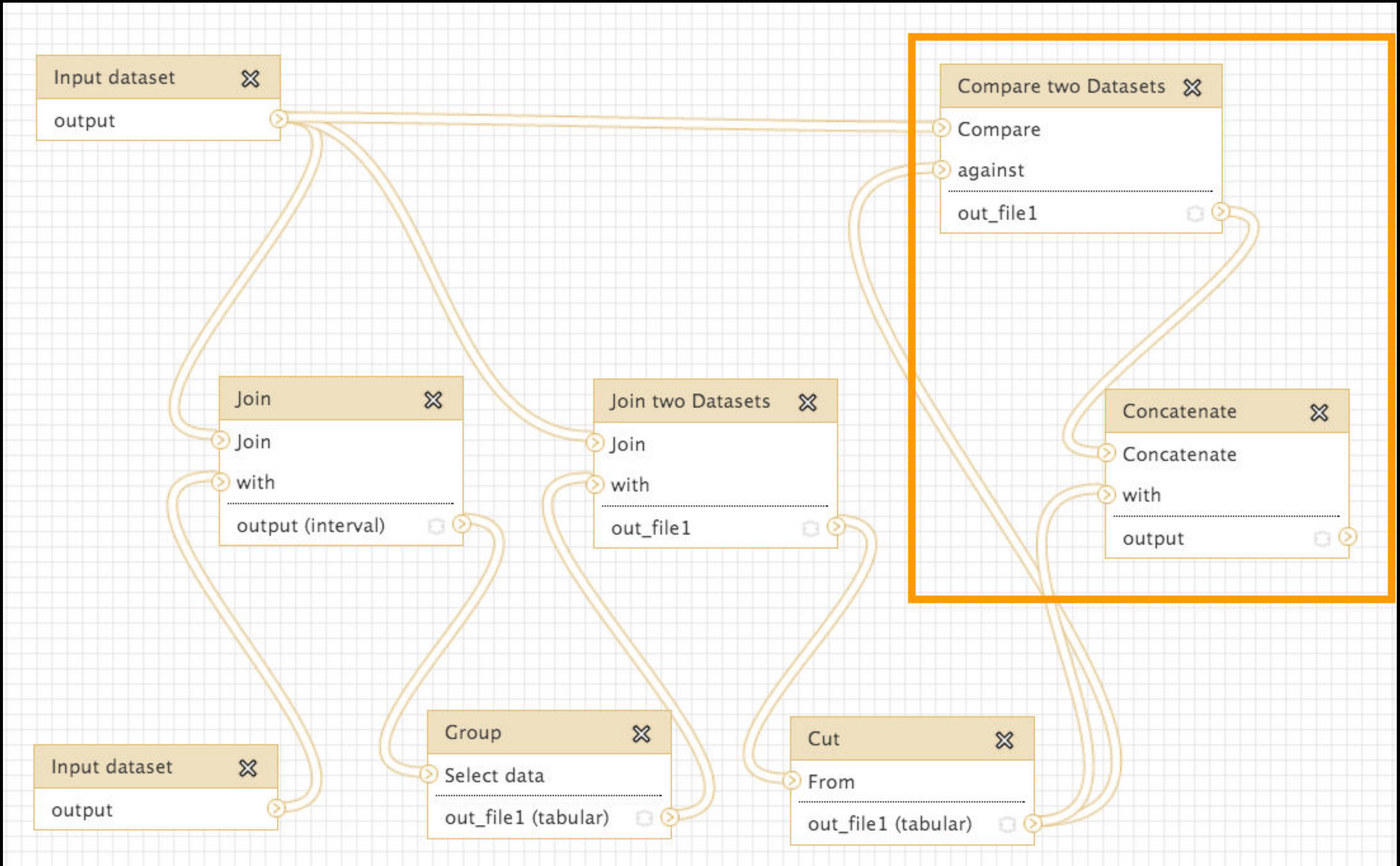
101: Overlapping Exons and Repeats

3.5 MB

Dataset	Annotation
1: Exons, chr22	
2: Repeats, chr22	
3: Join on data 2 and data 1	
4: Group on data 3	
5: Join two Datasets on data 1 and data 4	
6: Exons with overlapping repeats	

Note: In your solution, you can take advantage of the fact that Exons already have 0 scores.

One Possible Solution



Solution from Stanford Kwenda and Caron Griffiths, Pretoria.

Takes advantage of the fact that Exons already have 0 scores.

Basic Analysis: Further reading & Resources

http://usegalaxy.org/galaxy101

https://vimeo.com/76343659

Agenda

10:00	Welcome: Introduction and Logistics
10:15	Basic analysis with Galaxy
11:40	Galaxy Project Resources
12:00	Lunch (catered)
1:00	Advanced Usage: RNA-Seq Analysis
3:00	Done

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/

Galaxy Community Resources: Mailing Lists

http://wiki.galaxyproject.org/MailingLists

Galaxy-Dev

Questions about developing for and deploying Galaxy

High volume (5200 posts in 2013, 900+ members)
(3246 posts in 2014, 1000+ members)

Galaxy-Announce

Project announcements, low volume, moderated

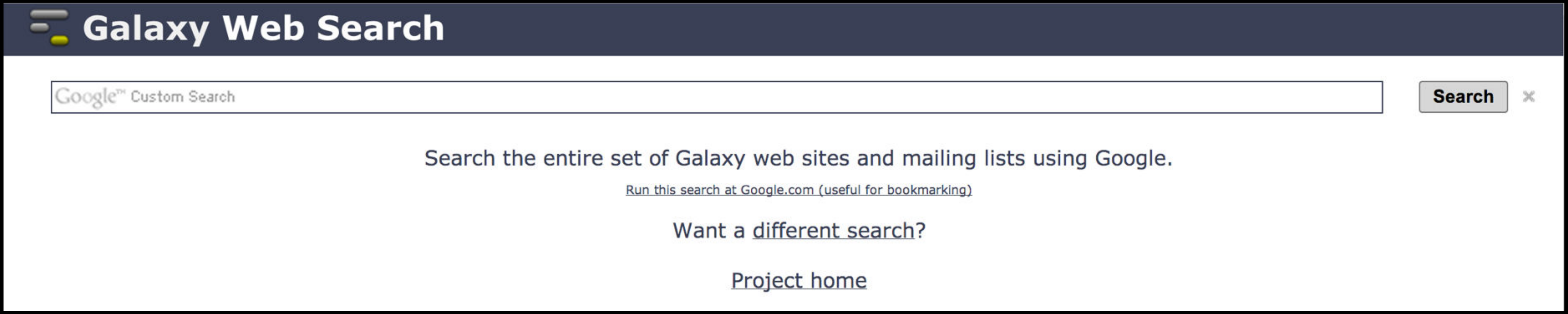
Low volume (47 posts in 2013, 3400+ members)
(34 posts in 2014, 4400+ members)

Galaxy-User (discontinued 2014/05)

Questions about using Galaxy and usegalaxy.org

High volume (1328 posts in 2013, 2600+ members)
(358 posts in 2014, 2600+ members)

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



Galaxy Web Search

Google™ Custom Search

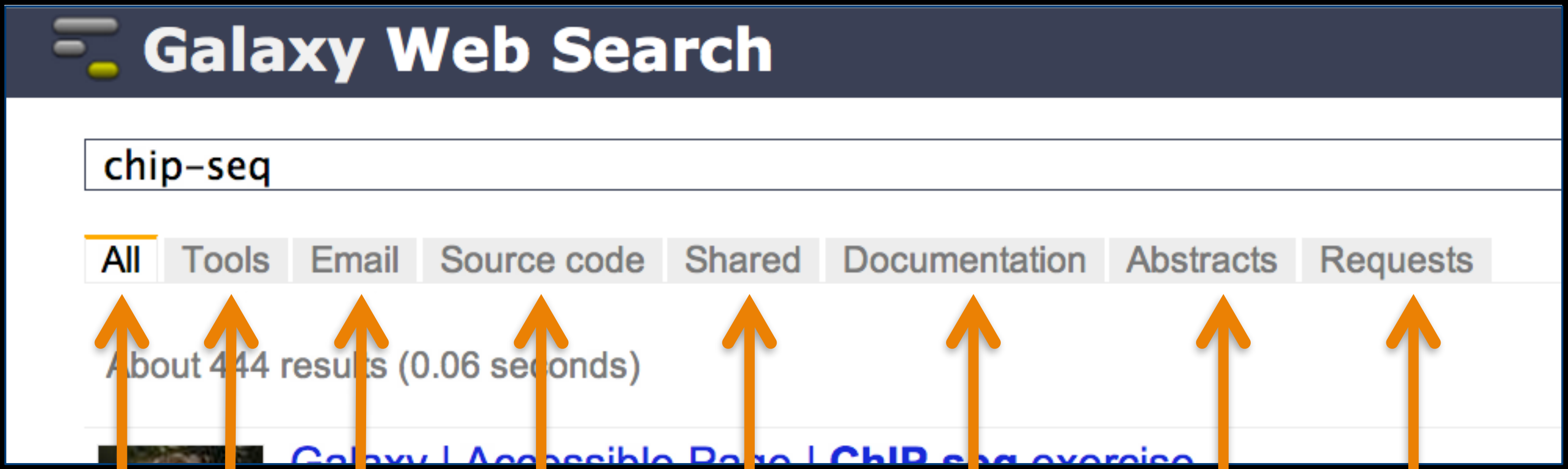
Search ×

Search the entire set of Galaxy web sites and mailing lists using Google.

Run this search at Google.com (useful for bookmarking)

Want a different search?

Project home



Galaxy Web Search

chip-seq

All | Tools | Email | Source code | Shared | Documentation | Abstracts | Requests

About 444 results (0.06 seconds)

Galaxy | Accessible Page | ChIP-seq exercise

Find

Everything on …

Tools for …

Email about …

Source code for …

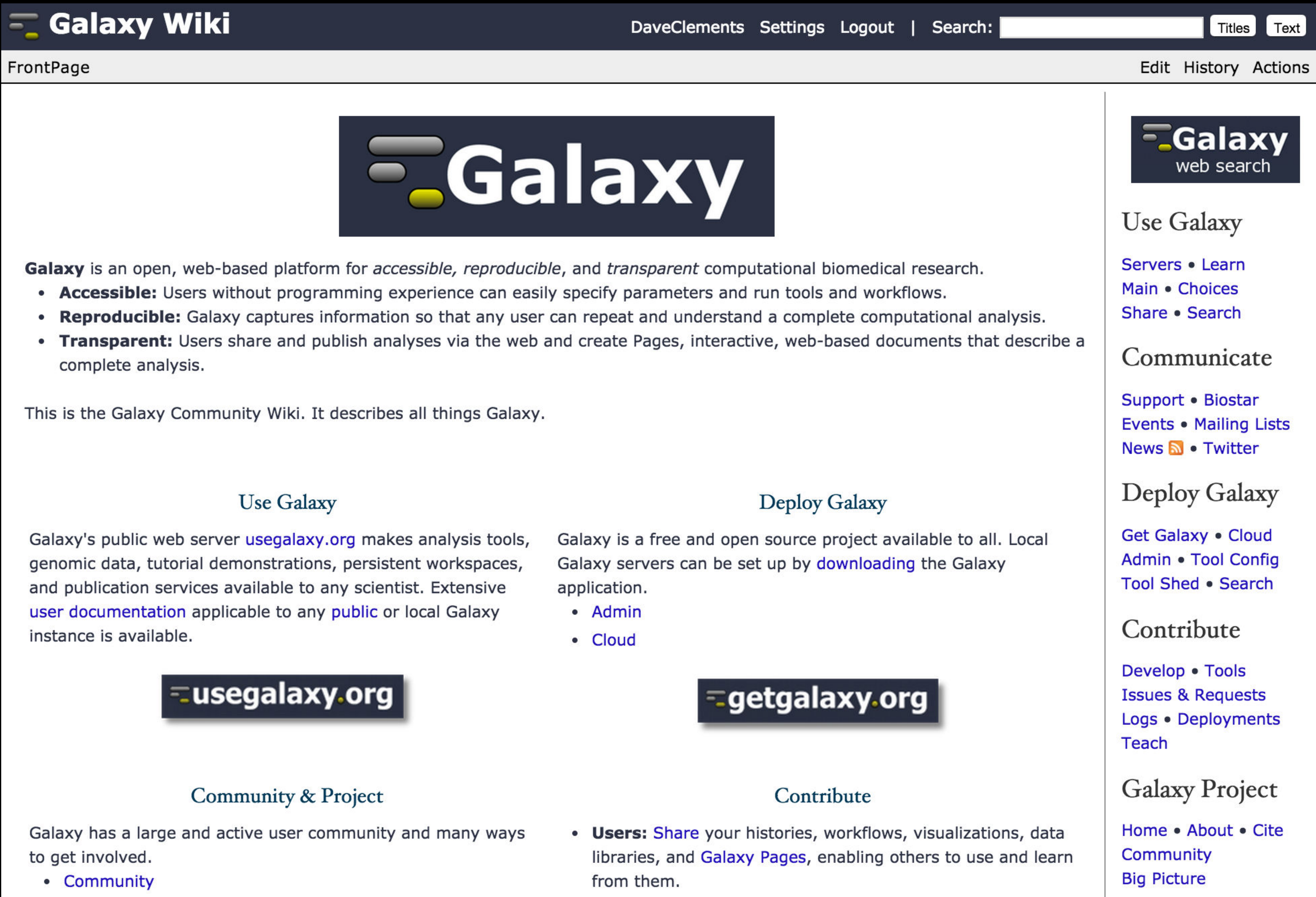
Published Histories, Pages, Workflows, about …

Documentation on …

Papers using Galaxy for …

Related feature requests

http://wiki.galaxyproject.org



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.

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

Community & Project

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

- Community

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

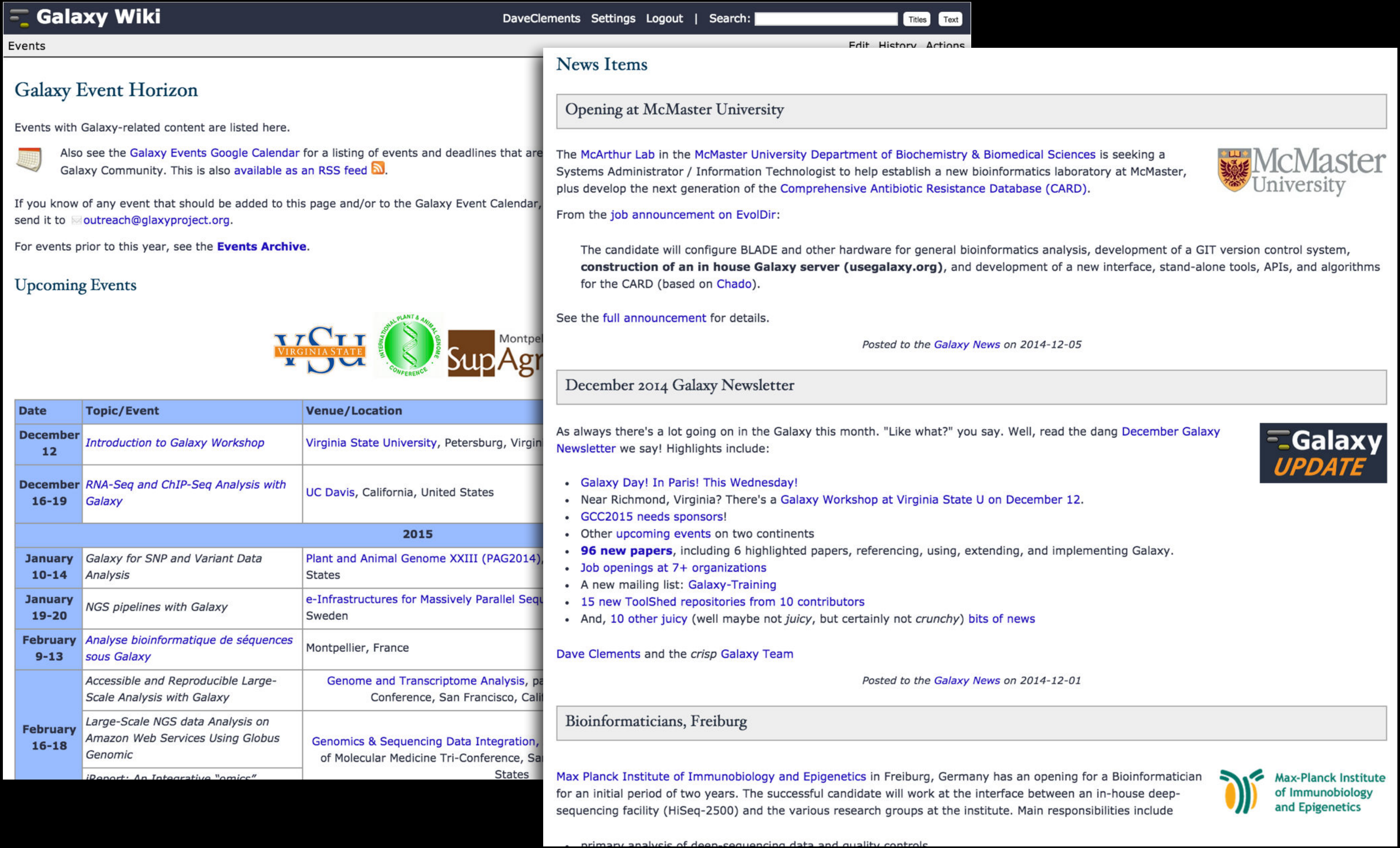
Develop • Tools
Issues & Requests
Logs • Deployments
Teach

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
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, Virgin
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 Sequ 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, pa Conference, San Francisco, Cali
	Large-Scale NGS data Analysis on Amazon Web Services Using Globus Genomic	Genomics & Sequencing Data Integration, of Molecular Medicine Tri-Conference, Sa States
	iReport: An Integrative "omics"	

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

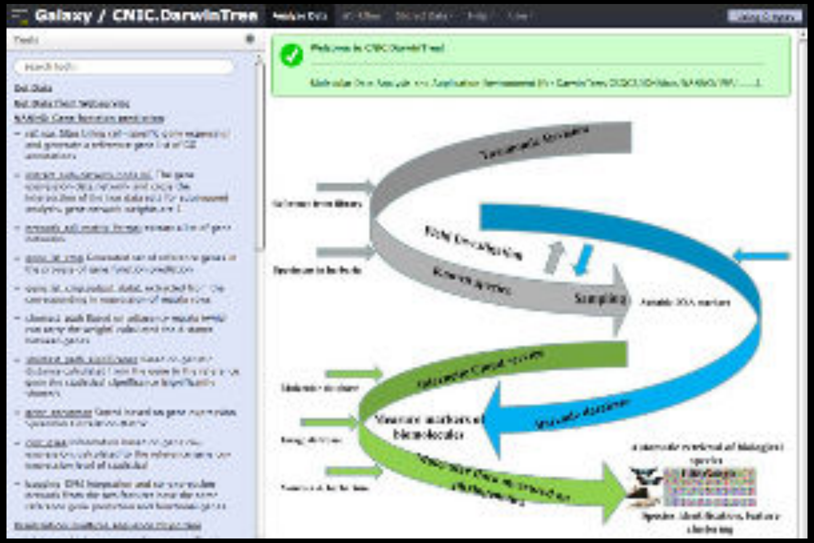
Bioinformaticians, Freiburg

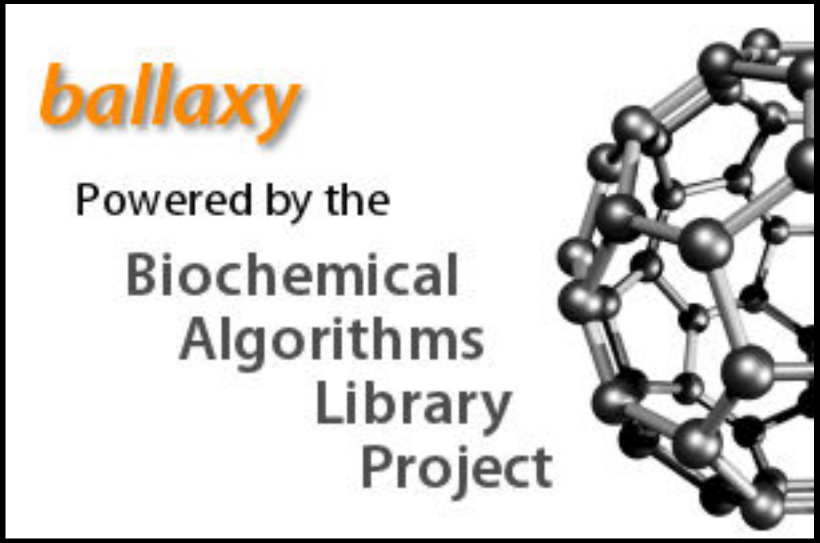
Max Planck Institute of Immunobiology and Epigenetics in Freiburg, Germany has an opening for a Bioinformatician for an initial period of two years. The successful candidate will work at the interface between an in-house deep-sequencing facility (HiSeq-2500) and the various research groups at the institute.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

- primary analysis of deep-sequencing data and quality controls

Cistrome

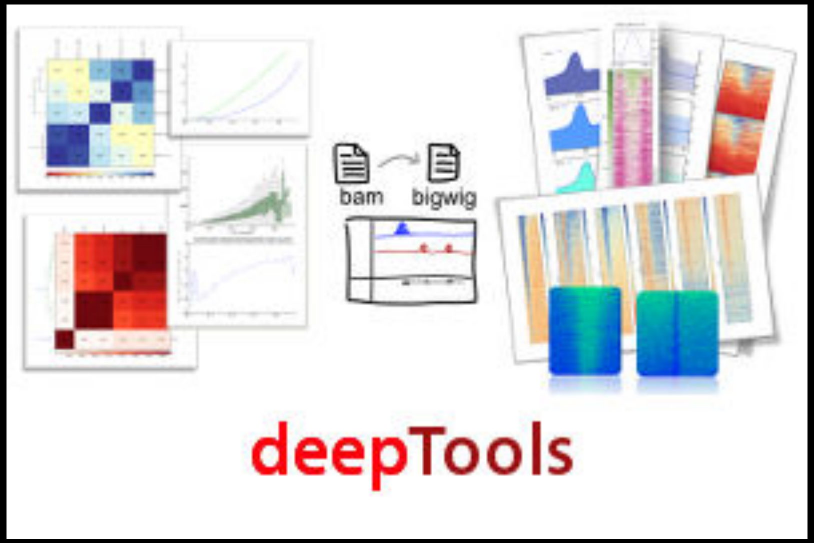
A Galaxy Server dedicated to ChIP-* analysis





ballaxy

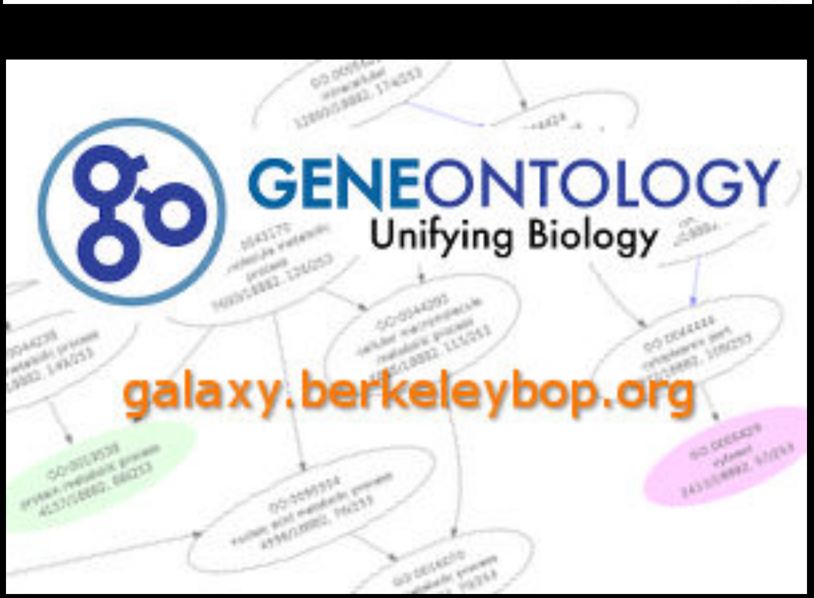
Powered by the Biochemical Algorithms Library Project



deepTools

CoSSci

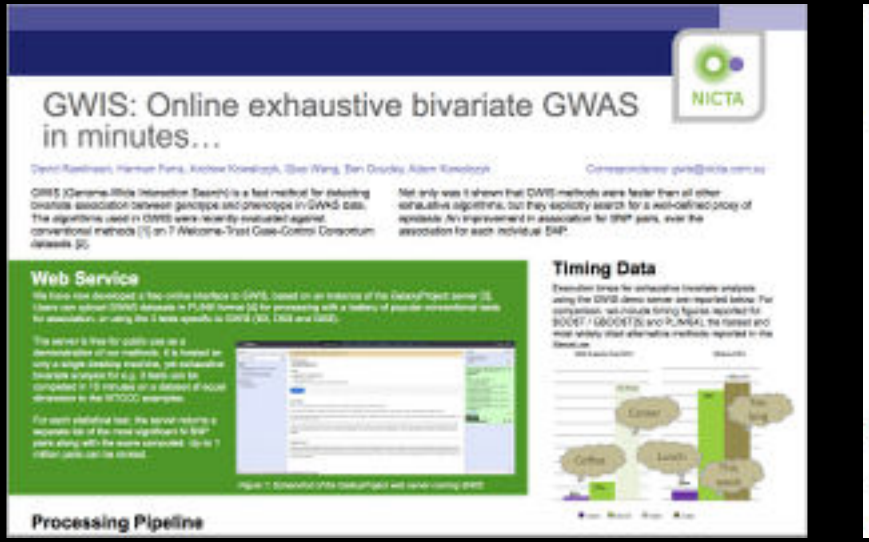
Galaxy for Complex Social Sciences



GENEONTOLOGY

Unifying Biology

galaxy.berkeleybop.org

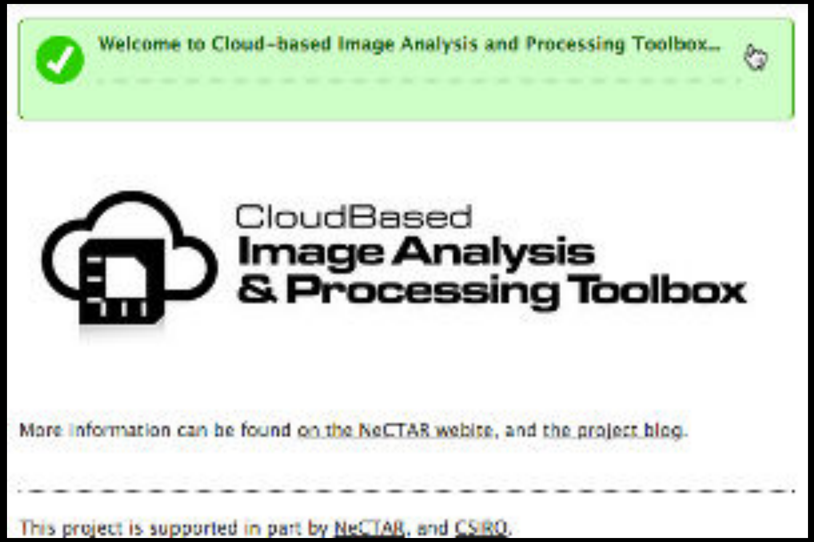


GWIS: Online exhaustive bivariate GWAS in minutes…

Web Service

Timing Data

Processing Pipeline

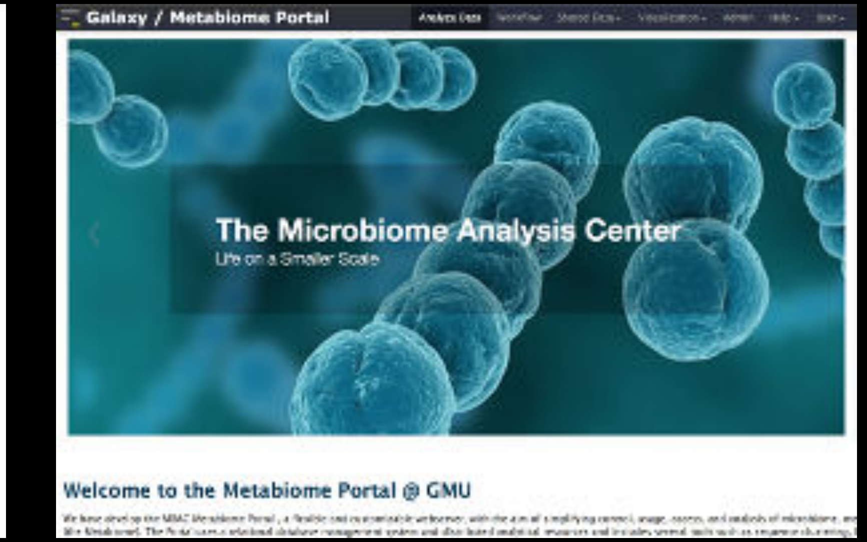


Welcome to Cloud-based Image Analysis and Processing Toolbox...

CloudBased Image Analysis & Processing Toolbox

More information can be found on the NeCTAR webite, and the project blog.

This project is supported in part by NeCTAR, and CSIRO.



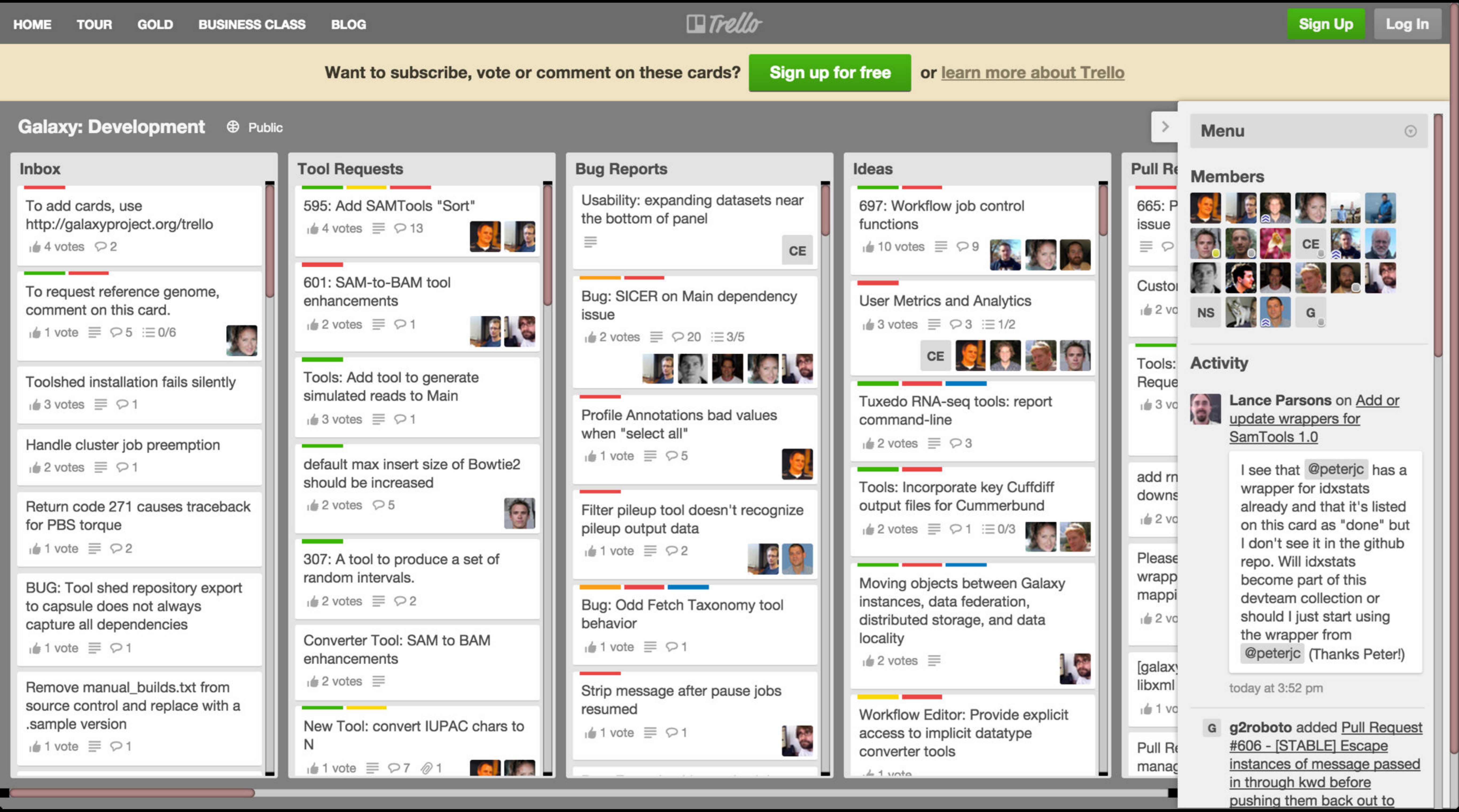
The Microbiome Analysis Center

Life on a Smaller Scale

Welcome to the Metabiome Portal @ GMU

bit.ly/gxyServers

Community can create, vote and comment on issues



Want to subscribe, vote or comment on these cards? Sign up for free or learn more about Trello

Galaxy: Development Public

Inbox

- To add cards, use http://galaxyproject.org/trello — 4 votes, 2
- To request reference genome, comment on this card. — 1 vote, 5, 0/6
- Toolshed installation fails silently — 3 votes, 1
- Handle cluster job preemption — 2 votes, 1
- Return code 271 causes traceback for PBS torque — 1 vote, 2
- BUG: Tool shed repository export to capsule does not always capture all dependencies — 1 vote, 1
- Remove manual_builds.txt from source control and replace with a .sample version — 1 vote, 1

Tool Requests

- 595: Add SAMTools "Sort" — 4 votes, 13
- 601: SAM-to-BAM tool enhancements — 2 votes, 1
- Tools: Add tool to generate simulated reads to Main — 3 votes, 1
- default max insert size of Bowtie2 should be increased — 2 votes, 5
- 307: A tool to produce a set of random intervals. — 2 votes, 2
- Converter Tool: SAM to BAM enhancements — 2 votes
- New Tool: convert IUPAC chars to N — 1 vote, 7, 1

Bug Reports

- Usability: expanding datasets near the bottom of panel — CE
- Bug: SICER on Main dependency issue — 2 votes, 20, 3/5
- Profile Annotations bad values when "select all" — 1 vote, 5
- Filter pileup tool doesn't recognize pileup output data — 1 vote, 2
- Bug: Odd Fetch Taxonomy tool behavior — 1 vote, 1
- Strip message after pause jobs resumed — 1 vote, 1

Ideas

- 697: Workflow job control functions — 10 votes, 9
- User Metrics and Analytics — 3 votes, 3, 1/2 — CE
- Tuxedo RNA-seq tools: report command-line — 2 votes, 3
- Tools: Incorporate key Cuffdiff output files for Cummerbund — 2 votes, 1, 0/3
- Moving objects between Galaxy instances, data federation, distributed storage, and data locality — 2 votes
- Workflow Editor: Provide explicit access to implicit datatype converter tools — 1 vote

Pull Re

- 665: P issue
- Custo — 2 vo
- Tools: Reque — 3 vo
- add rn downs — 2 vo
- Please wrapp mappi — 2 vo
- [galax libxml — 1 vo
- Pull Re manag

Menu

Members

CE NS G

Activity

Lance Parsons on Add or update wrappers for SamTools 1.0

I see that @peterjc has a wrapper for idxstats already and that it's listed on this card as "done" but I don't see it in the github repo. Will idxstats become part of this devteam collection or should I just start using the wrapper from @peterjc (Thanks Peter!)

today at 3:52 pm

g2roboto added Pull Request #606 - [STABLE] Escape instances of message passed in through kwd before pushing them back out to

http://bit.ly/gxytrello

GALAXY COMMUNITY CONFERENCE

BALTIMORE, MD | JUNE 30 - JULY 2, 2014

Slides, posters & videos now online
http://bit.ly/gcc2014

Galaxy
Community
Conference
2012
July
25-27
UIC
Forum
University of Illinois at Chicago
http://galaxyproject.org/GCC2012
UIC
Ci
I

GCC 2015

Galaxy Community Conference

6-8th July 2015

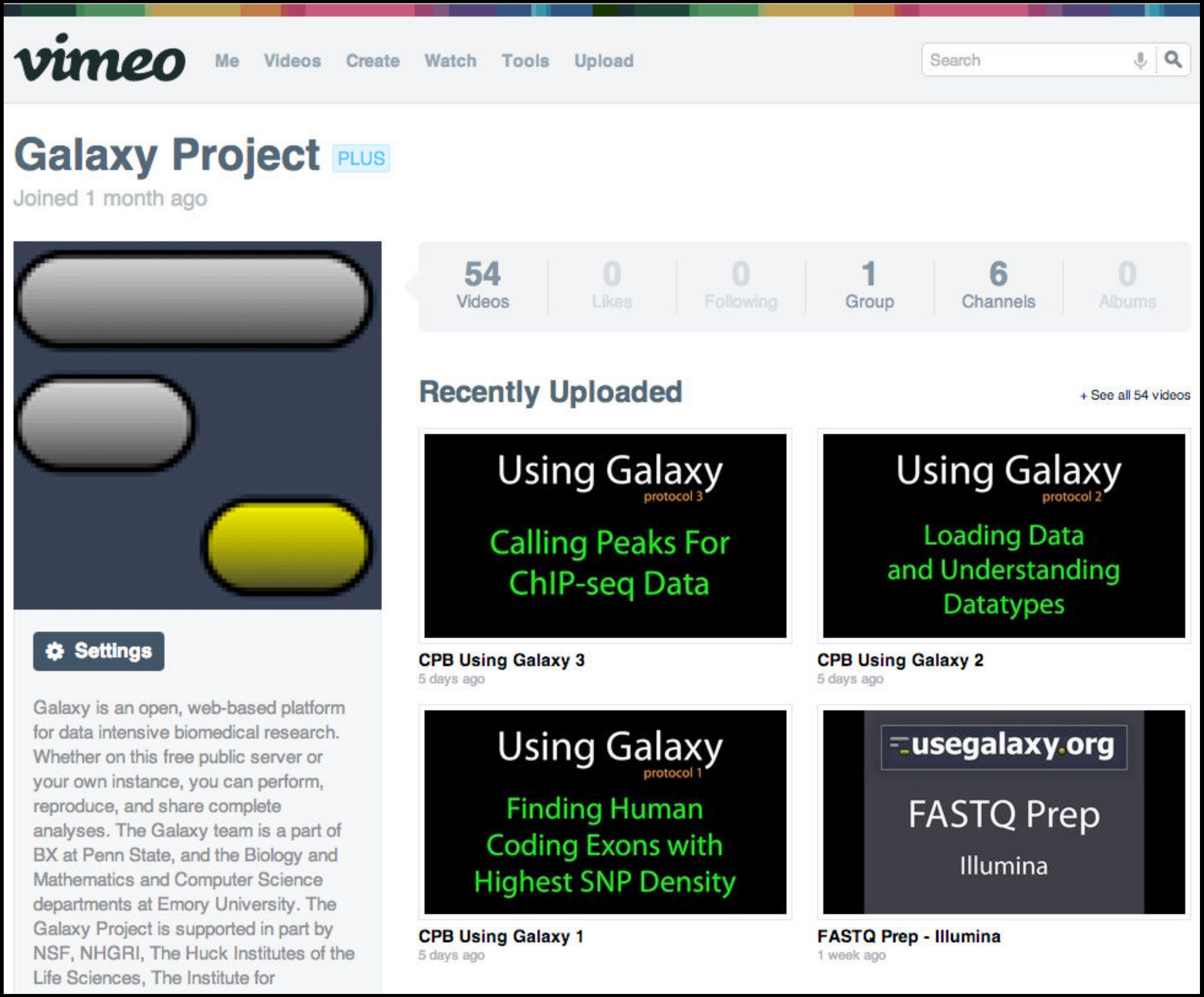
The Sainsbury Laboratory
Norwich, UK

galaxyproject.org

We also support community organized efforts and events.

swiss german galaxy tour
Bern
30 Sep - 1 Oct
Freiburg
2 Oct

Galaxy Resources & Community: Videos



Galaxy Project PLUS

Joined 1 month ago

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

⚙ 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

Recently Uploaded

+ See all 54 videos

CPB Using Galaxy 3

5 days ago

CPB Using Galaxy 2

5 days ago

CPB Using Galaxy 1

5 days ago

FASTQ Prep - Illumina

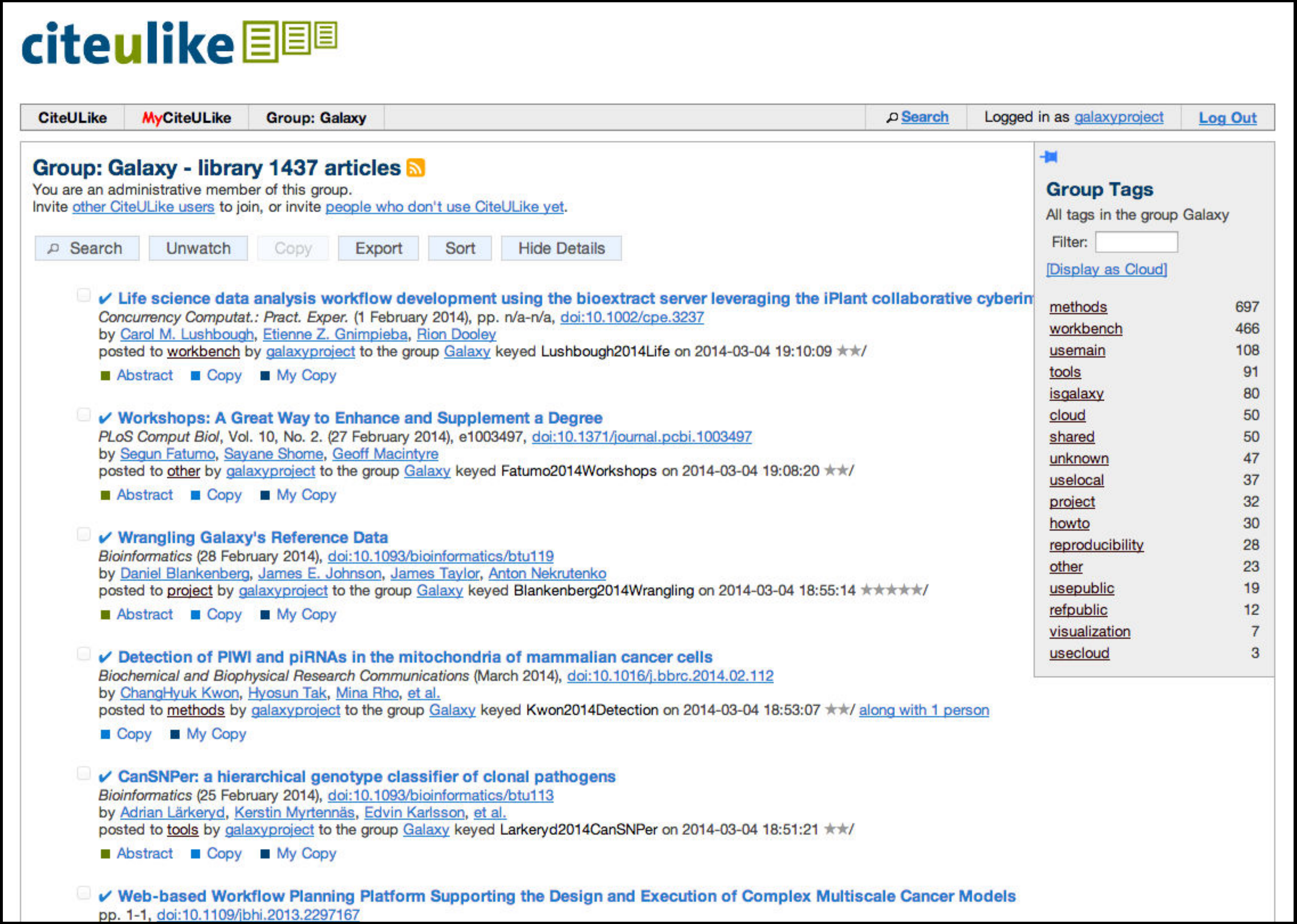
1 week ago

"How to" screencasts on using and deploying Galaxy

Talks from previous meetings.

http://vimeo.com/galaxyproject

Galaxy Resources & Community: CiteULike Group



citeulike

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

Group: Galaxy - library 1437 articles

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

□ ✓ **Life science data analysis workflow development using the bioextract server leveraging the iPlant collaborative cyberin**
Concurrency Computat.: Pract. Exper. (1 February 2014), pp. n/a-n/a, doi:10.1002/cpe.3237
by Carol M. Lushbough, Etienne Z. Gnimpieba, Rion Dooley
posted to workbench by galaxyproject to the group Galaxy keyed Lushbough2014Life on 2014-03-04 19:10:09 ★★/
Abstract | Copy | My Copy

□ ✓ **Workshops: A Great Way to Enhance and Supplement a Degree**
PLoS Comput Biol, Vol. 10, No. 2. (27 February 2014), e1003497, doi:10.1371/journal.pcbi.1003497
by Segun Fatumo, Sayane Shome, Geoff Macintyre
posted to other by galaxyproject to the group Galaxy keyed Fatumo2014Workshops on 2014-03-04 19:08:20 ★★/
Abstract | Copy | My Copy

□ ✓ **Wrangling Galaxy's Reference Data**
Bioinformatics (28 February 2014), doi:10.1093/bioinformatics/btu119
by Daniel Blankenberg, James E. Johnson, James Taylor, Anton Nekrutenko
posted to project by galaxyproject to the group Galaxy keyed Blankenberg2014Wrangling on 2014-03-04 18:55:14 ★★★★★/
Abstract | Copy | My Copy

□ ✓ **Detection of PIWI and piRNAs in the mitochondria of mammalian cancer cells**
Biochemical and Biophysical Research Communications (March 2014), doi:10.1016/j.bbrc.2014.02.112
by ChangHyuk Kwon, Hyosun Tak, Mina Rho, et al.
posted to methods by galaxyproject to the group Galaxy keyed Kwon2014Detection on 2014-03-04 18:53:07 ★★/ along with 1 person
Copy | My Copy

□ ✓ **CanSNPer: a hierarchical genotype classifier of clonal pathogens**
Bioinformatics (25 February 2014), doi:10.1093/bioinformatics/btu113
by Adrian Lärkeryd, Kerstin Myrtennäs, Edvin Karlsson, et al.
posted to tools by galaxyproject to the group Galaxy keyed Larkeryd2014CanSNPer on 2014-03-04 18:51:21 ★★/
Abstract | Copy | My Copy

□ ✓ **Web-based Workflow Planning Platform Supporting the Design and Execution of Complex Multiscale Cancer Models**
pp. 1-1, doi:10.1109/jbhi.2013.2297167

Group Tags

All tags in the group Galaxy

Filter:

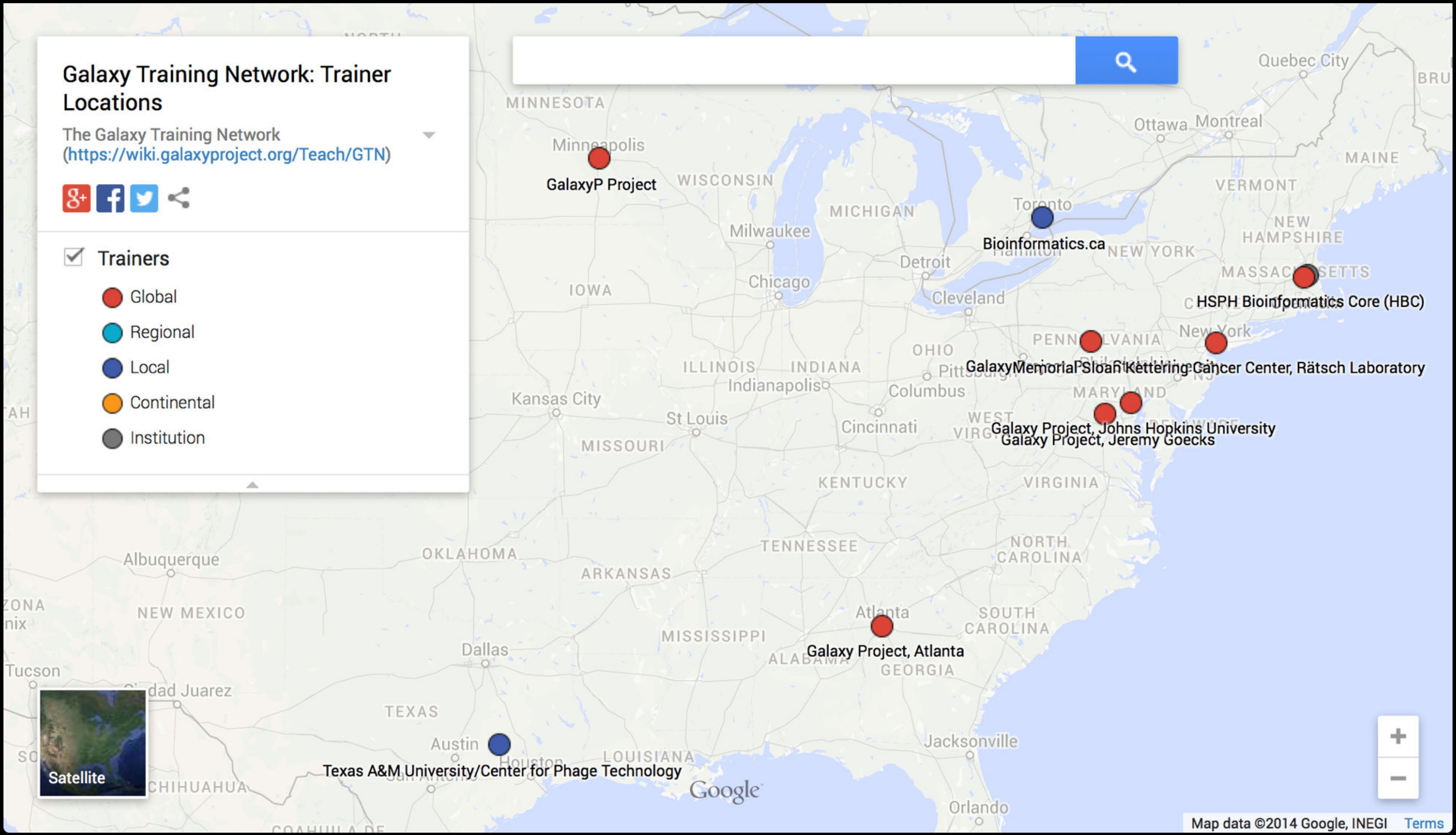
[Display as Cloud]

methods	697
workbench	466
usemain	108
tools	91
isgalaxy	80
cloud	50
shared	50
unknown	47
uselocal	37
project	32
howto	30
reproducibility	28
other	23
usepublic	19
refpublic	12
visualization	7
usecloud	3

Over 1900 papers

http://bit.ly/gxycul

Scaling the Project: Training



Galaxy Training Network: Trainer Locations

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

☑ Trainers

- ● Global
- ● Regional
- ● Local
- ● Continental
- ● Institution

GalaxyP Project
Bioinformatics.ca
HSPH Bioinformatics Core (HBC)
GalaxyMemorial Sloan Kettering Cancer Center, Rätsch Laboratory
Galaxy Project, Johns Hopkins University
Galaxy Project, Jeremy Goecks
Galaxy Project, Atlanta
Texas A&M University/Center for Phage Technology
Satellite
Google
Map data ©2014 Google, INEGI
Terms

Galaxy Training Network launched In October.

bit.ly/gxygtn

The Galaxy Team

Enis Afgan

Dannon Baker

Dan Blankenberg

Dave Bouvier

Marten Cech

John Chilton

Dave Clements

Nate Coraor

Carl Eberhard

Jeremy Goecks

Sam Guerler

Jen Jackson

Ross Lazarus

Anton Nekrutenko

Nick Stoler

James Taylor

Nitesh Turaga

http://wiki.galaxyproject.org/GalaxyTeam

Galaxy is hiring post-docs and software engineers

Please help.

http://wiki.galaxyproject.org/GalaxyIsHiring

Also Thanks To

VSU VIRGINIA STATE

College of Natural and Health Sciences

Glenn Harris

National Institutes of Health
Amazon Web Services

Thanks

Dave Clements

Galaxy Project

Johns Hopkins University

outreach@galaxyproject.org

Agenda

10:00 Welcome:
Introduction and Logistics

10:15 Basic analysis with Galaxy

11:40 Galaxy Project Resources

12:00 Lunch (catered)

1:00 Advanced Usage: RNA-Seq Analysis

3:00 Done

Agenda

10:00	Welcome: Introduction and Logistics
10:15	Basic analysis with Galaxy
11:40	Galaxy Project Resources
12:00	Lunch (catered)
1:00	Advanced Usage: RNA-Seq Analysis
3:00	Done

RNA-Seq Analysis: Get the Data

Create new history

⚙ (cog) → Create New

Import:

Shared Data → Data Libraries

→ RNA-Seq UCDavis 2013 Example Data*

→ Unfiltered Reads

→ MeOH_REP1_R1.fastq and MeOH_REP1_R2.fastq

UCDAVIS Bioinformatics Core
Genome Center

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

NGS Data Quality Control

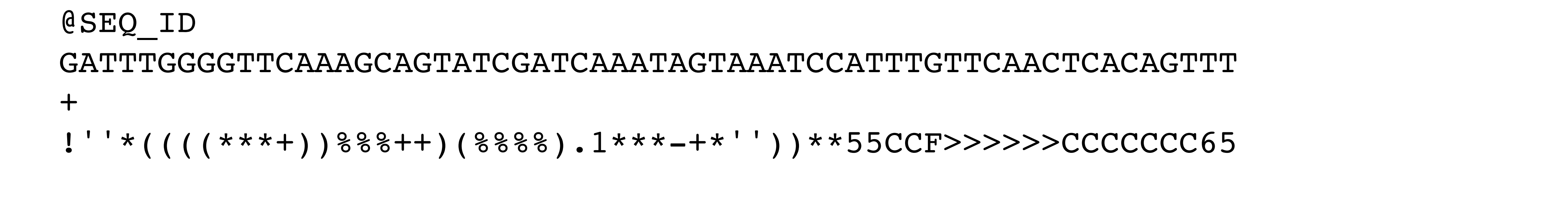
- FASTQ format
- Examine quality in an RNA-Seq dataset
- Trim/filter as we see fit, hopefully without breaking anything.

Quality Control is not sexy.

It is vital.

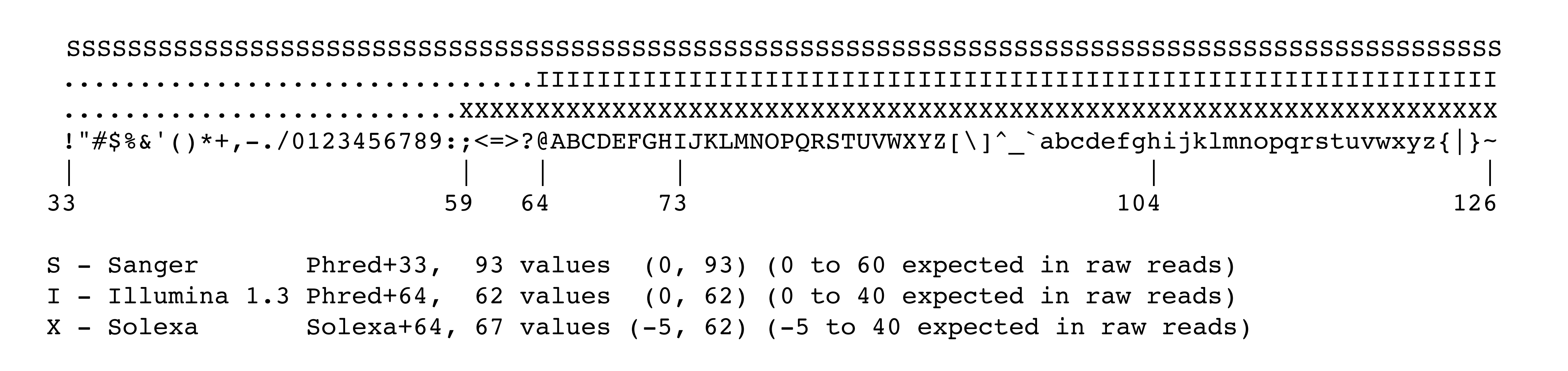
What is FASTQ?

- 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!



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126

S - Sanger       Phred+33,  93 values  (0, 93) (0 to 60 expected in raw reads)
I - Illumina 1.3 Phred+64,  62 values  (0, 62) (0 to 40 expected in raw reads)
X - Solexa       Solexa+64, 67 values (-5, 62) (-5 to 40 expected in raw reads)
```

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

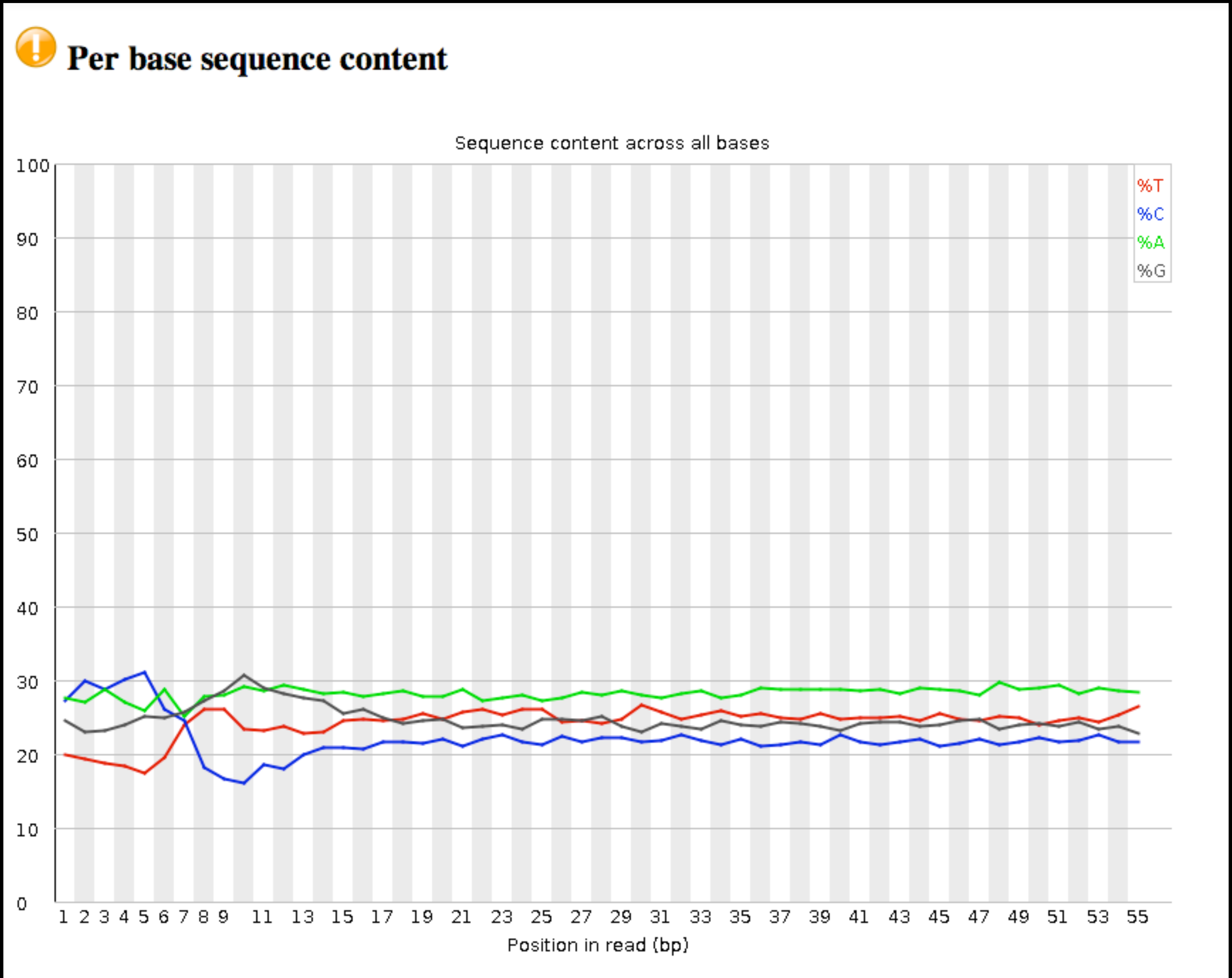
NGS Data Quality: Assessment tools

NGS QC and Manipulation → FastQC

Gives you a lot of information but little control over how it is calculated or presented.

http://bit.ly/FastQCBoxPlot

NGS Data Quality: Sequence bias at front of reads?

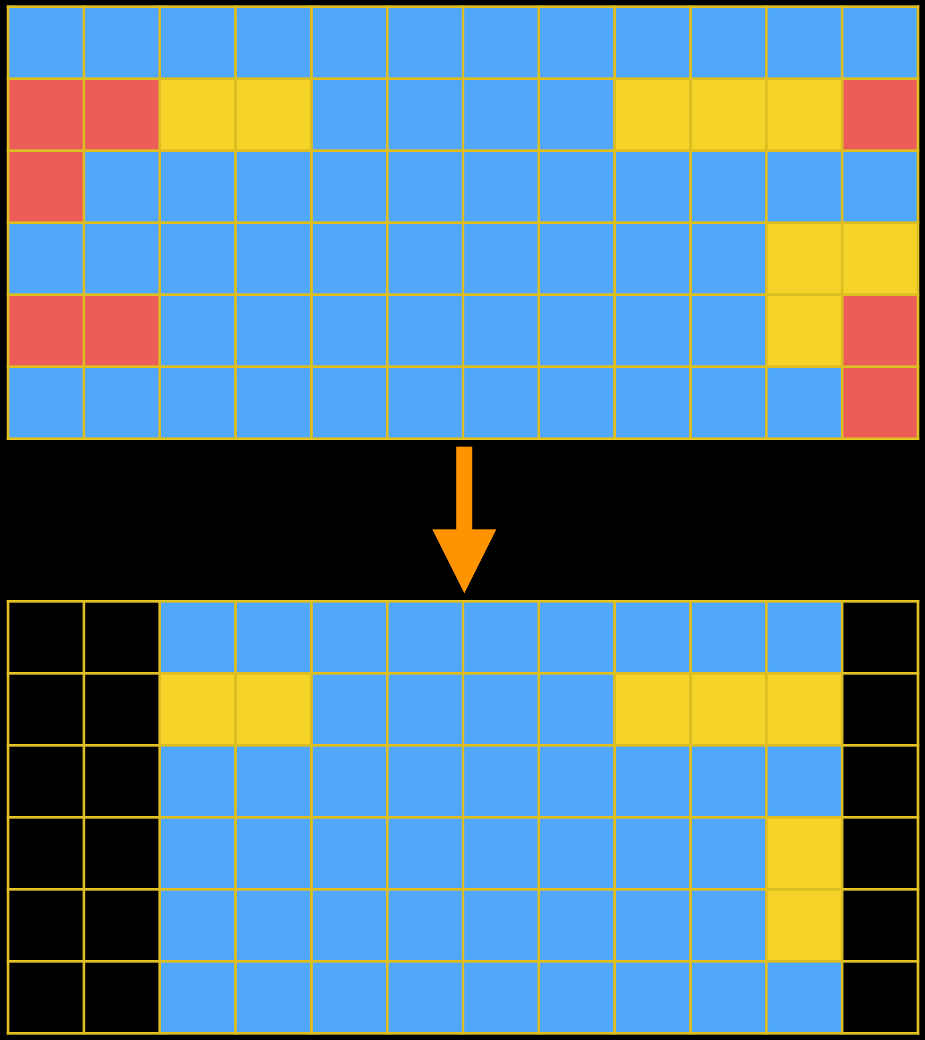


From a sequence specific bias that is caused by use of random hexamers in library preparation.

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

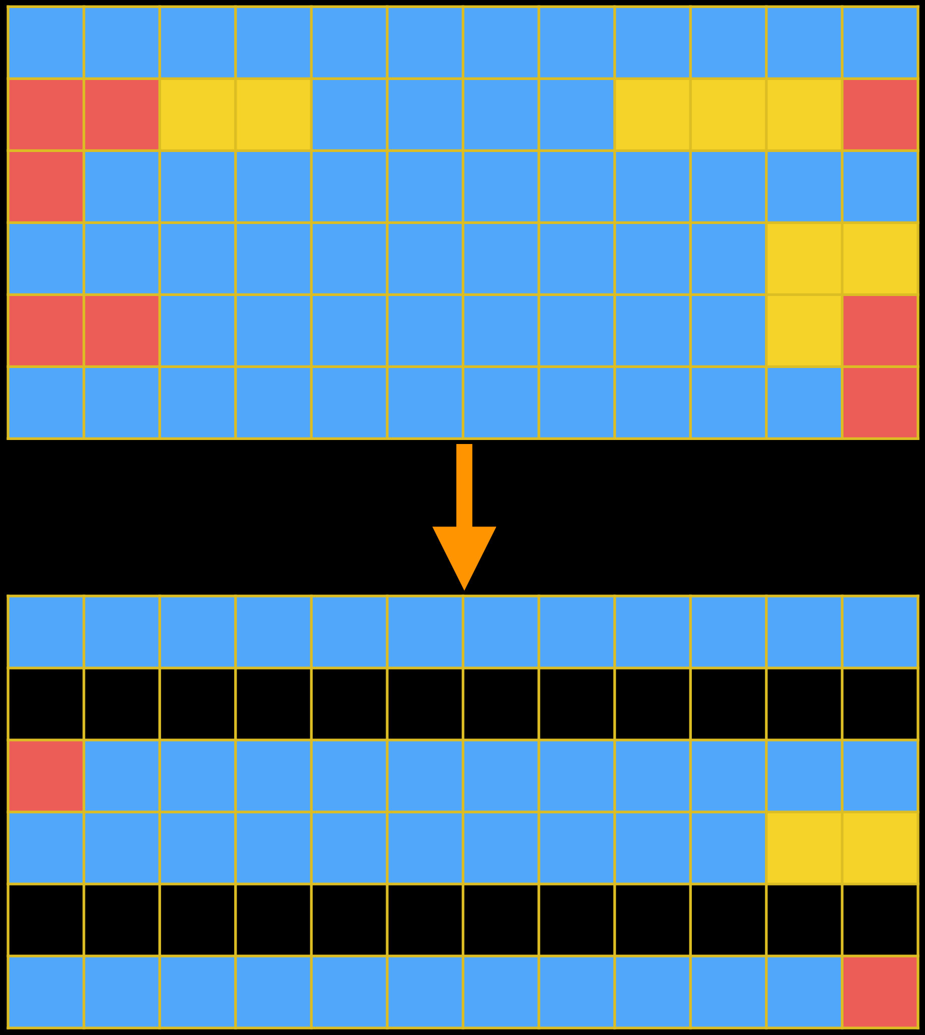
NGS Data Quality: Trim as we see fit

- Trim as we see fit: Option 1
  - **NGS QC and Manipulation → FASTQ Trimmer by column**
  - Trim same number of columns from every record
  - Can specify different trim for 5′ and 3′ ends



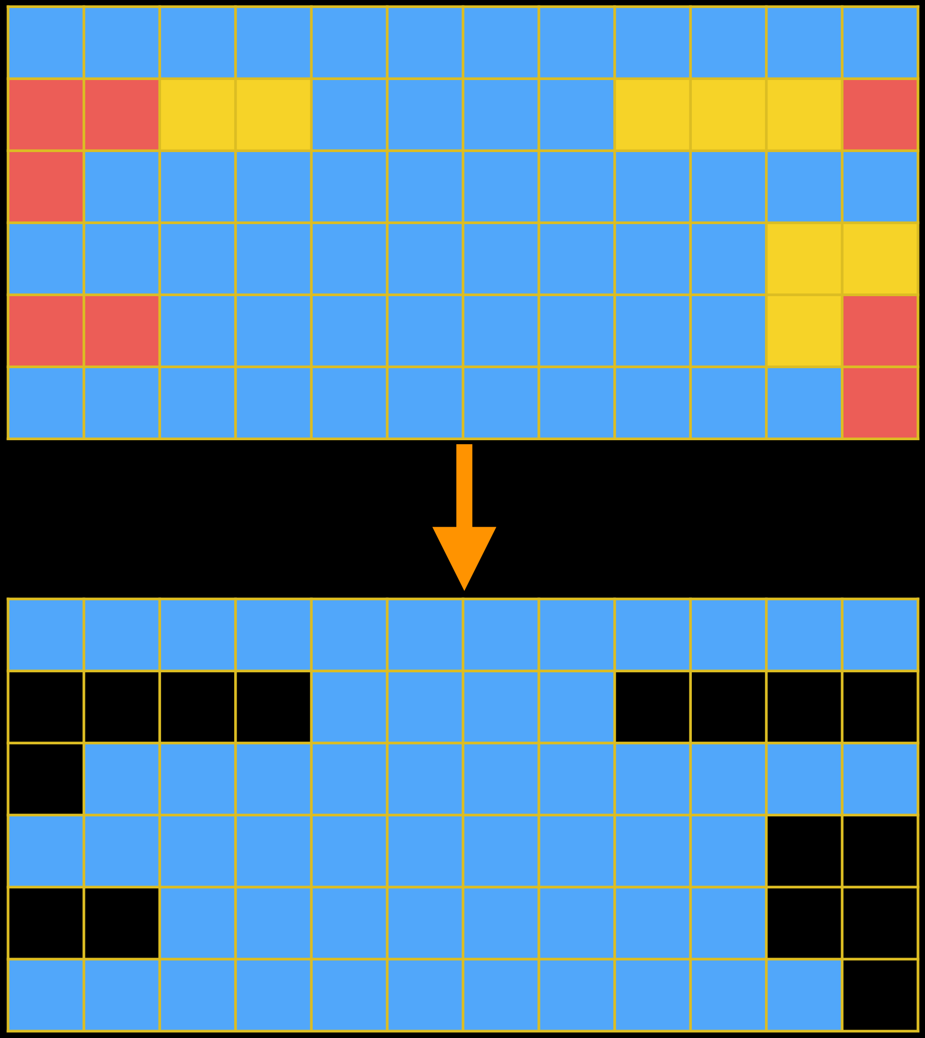
NGS Data Quality: Base Quality Trimming

- ~~Trim~~ Filter as we see fit: Option 2
  - NGS QC and Manipulation → **Filter FASTQ reads by quality score and length**
  - Keep or discard whole reads
  - Can have different thresholds for different regions of the reads.
  - Keeps original read length.

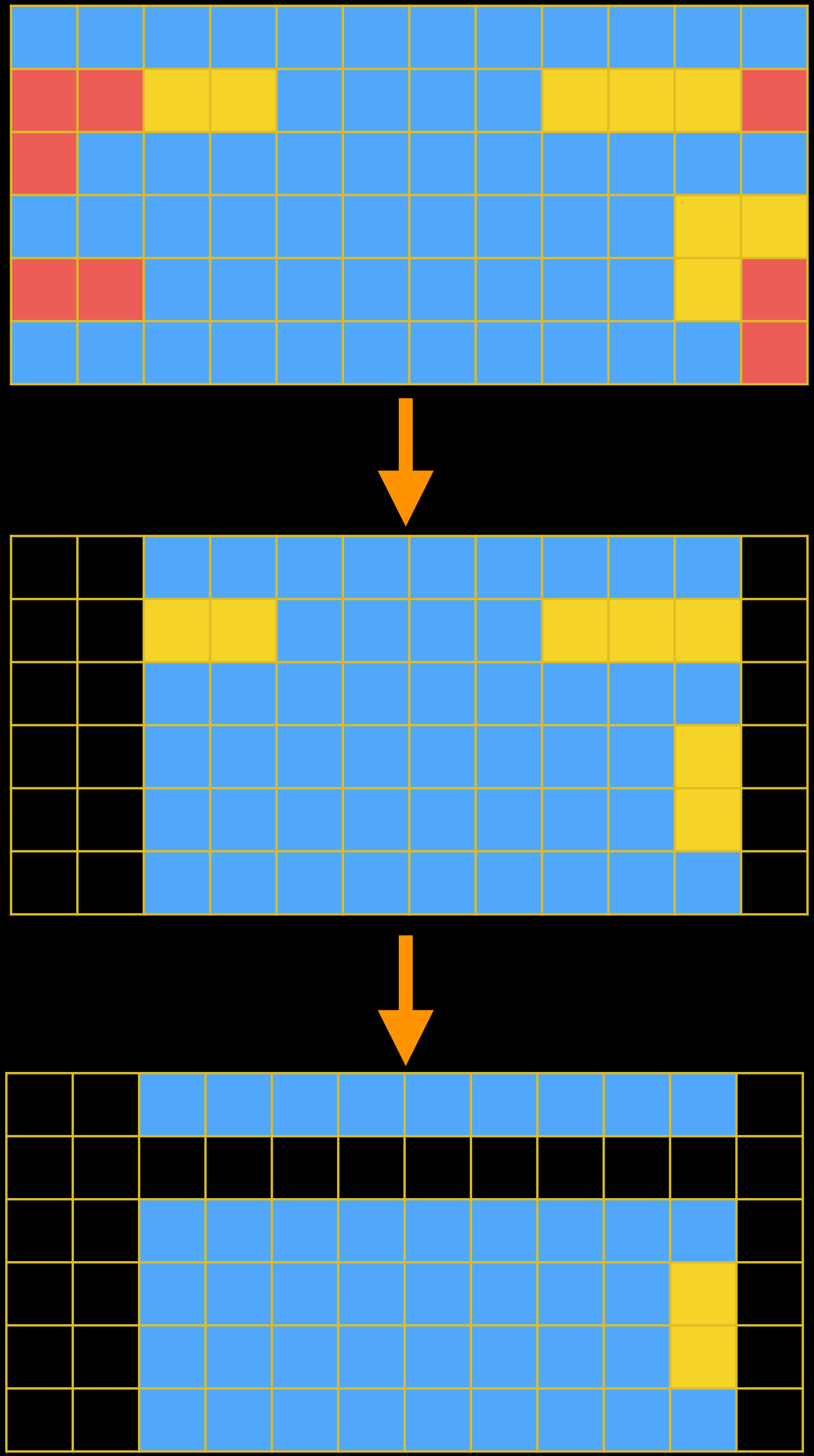


NGS Data Quality: Base Quality Trimming

- Trim as we see fit: Option 3
  - NGS QC and Manipulation → **FASTQ Quality Trimmer by sliding window**
  - Trim from both ends, using sliding windows, until you hit a high-quality section.
  - Produces variable length reads



Options are not mutually exclusive



Option 1

(by column)

+

Option 2

(by entire row)

Trim? *As we see fit?*

- Introduced 3 options
  - One preserves original read length, two don’t
  - One preserves number of reads, two don’t
  - Two keep/make every read the same length, one does not
  - One preserves pairings, two don't

Trim? *As we see fit?*

- Choice depends on downstream tools
- Find out assumptions & requirements for downstream tools and make appropriate choice(s) now.
- How to do that?
  - Read the tool documentation
  - http://biostars.org/
  - http://seqanswers.com/
  - http://galaxyproject.org/search

SEQanswers
the next generation sequencing community

Biostar
BIOINFORMATICS EXPLAINED

“Mixing paired- and single- end reads together is **not** supported.”

Tophat Manual

“If you are performing RNA-seq analysis, there is no need to filter the data to ensure exact pairs before running Tophat.”

Jen Jackson

Galaxy User Support Person Extraordinaire

“Dang.”

Most of us

Running Tophat on *no-longer-cleanly-paired* data *does map the reads*, but, it no longer keeps track of read pairs in the SAM/BAM file.

Keeping paired ends paired: Things to Try

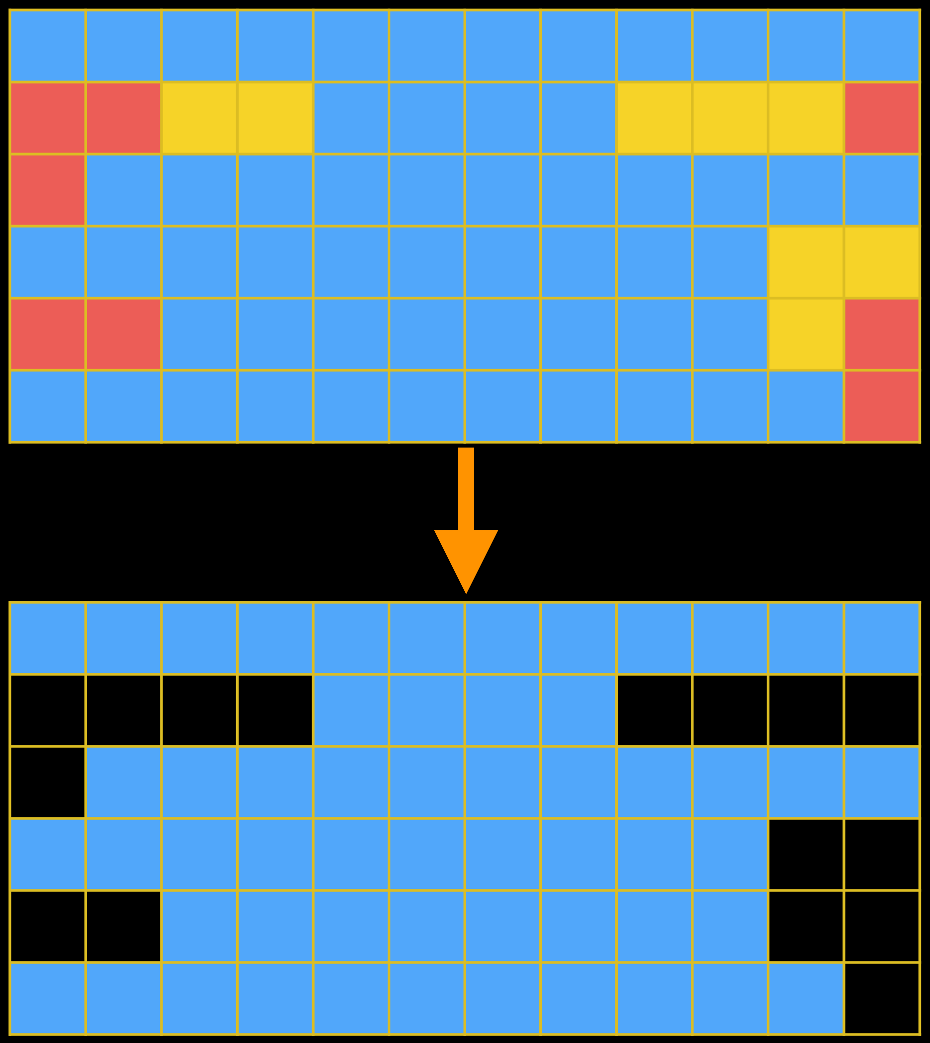
- Don't bother.

- Run a workflow (try the "Re-Pair Paired ends after QC may have broken them" workflow) that removes any unpaired reads before mapping:

- Run the Picard Paired Read Mate Fixer after mapping reads.

- Use sliding windows for QC, but keep empty reads. (This does not work with Tophat.)

NGS Data Quality: Base Quality Trimming



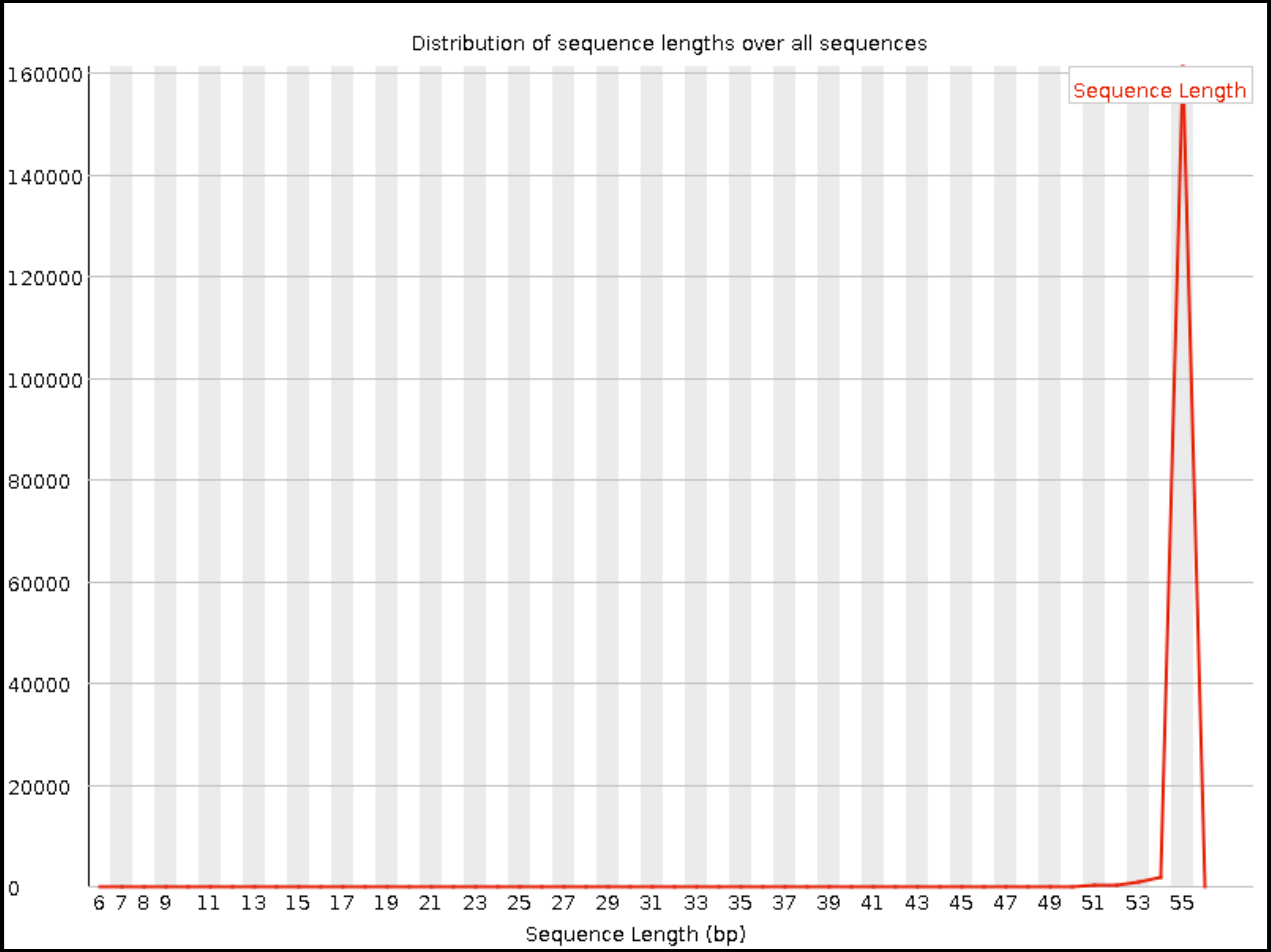
I'll use Option3, sliding windows, and run a workflow afterward to patch up pairings

- NGS QC and Manipulation → **FASTQ Quality Trimmer by sliding window**

Run again:

- NGS QC and Manipulation → **FastQC** on trimmed dataset

NGS Data Quality: Base Quality Trimming



New Problem?

Now some reads are so short they are just noise and can’t be meaningfully mapped. Have potential to bog down mapping.

Option 2 can fix this, but breaks pairings (if you still have them).

Or, your mapper may have an option to ignore shorter reads.

RNA-Seq Analysis

I'll use option 2, since my pairings are already broken.

NGS QC and Manipulation

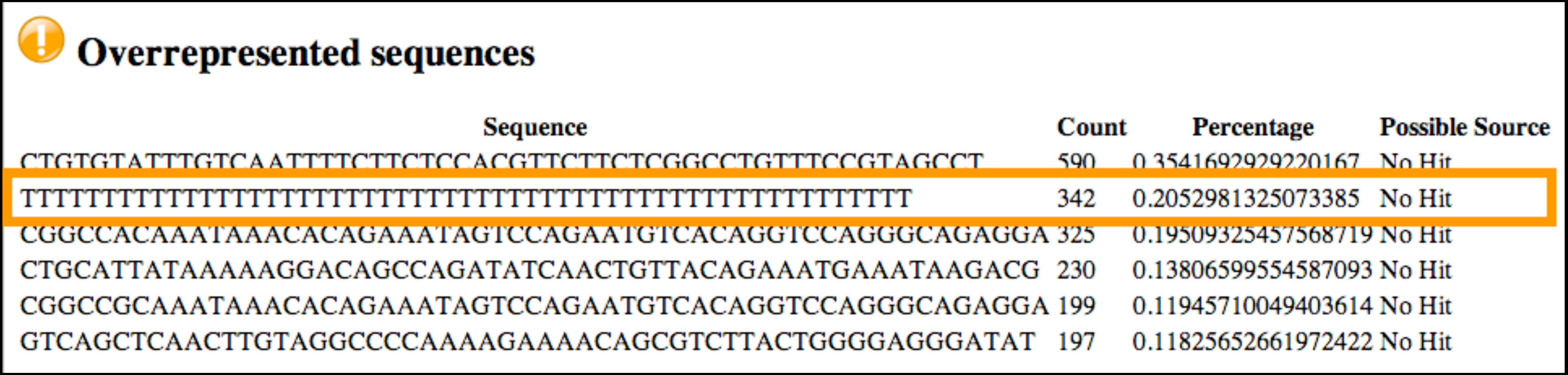
→ **Filter FASTQ reads by quality score and length**

Pick a minimum length. I used 32.

NGS Data Quality: Sequencing Artifacts

Repeat this process with MeOH Rep1 R2 (the reverse reads)

... and now we notice a problem in Overrepresented sequences:



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CTGTGTATTTGTCAATTTTCTTCTCCACGTTCTTCTCGGCCTGTTTCCGTAGCCT	590	0.3541692929220167	No Hit
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	342	0.2052981325073385	No Hit
CGGCCACAAATAAACACAGAAATAGTCCAGAATGTCACAGGTCCAGGGCAGAGGA	325	0.19509325457568719	No Hit
CTGCATTATAAAAAGGACAGCCAGATATCAACTGTTACAGAAATGAAATAAGACG	230	0.13806599554587093	No Hit
CGGCCGCAAATAAACACAGAAATAGTCCAGAATGTCACAGGTCCAGGGCAGAGGA	199	0.11945710049403614	No Hit
GTCAGCTCAACTTGTAGGCCCCAAAAGAAAACAGCGTCTTACTGGGGAGGGATAT	197	0.11825652661972422	No Hit

NGS QC and Manipulation → **Remove sequencing artifacts**

But this will break pairings (if we still have them).

Or, can rely on mapper to just not map them.

RNA-Seq Analysis: Restore Pairings

If your QC filters might have broken pairings, then you may want to restore them.

Shared Data → Published Workflows

→ Re-Pair Paired ends after QC may have broken them

→ Import

Workflows

→ Re-Pair Paired ends after QC may have broken them

→ Run

Re-Pair Paired ends after QC may have broken them

Workflow takes 4 inputs

- Forward Reads, before QC
- Reverse Reads, before QC
- Forward Reads, after QC
- Reverse Reads, after QC

And produces 4 outputs

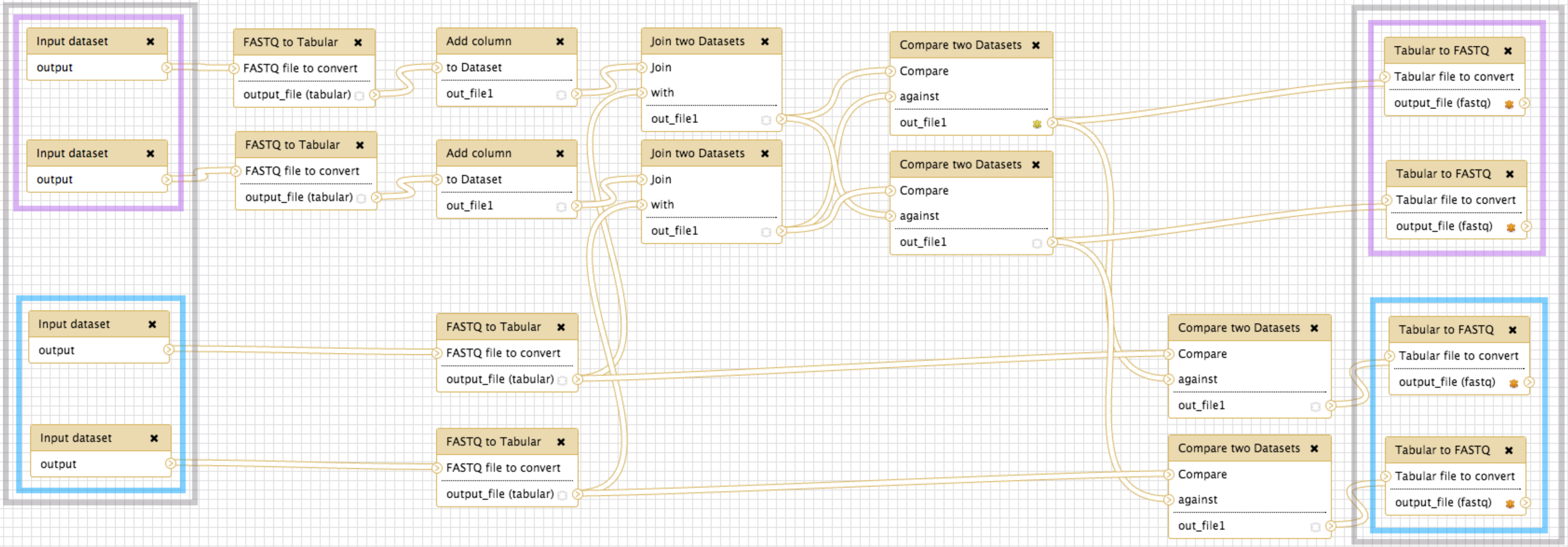
- Forward reads, re-paired
- Reverse reads, re-paired
- Forward reads, singletons
- Reverse reads, singletons

Workflow assumes pre-QC reads are correctly paired

Re-Pair Paired ends after QC may have broken them

Inputs

Outputs



□ Correctly Paired Reads

□ Incorrectly Paired / Unpaired Reads

NGS Data Quality: Done with 1st Replicate!

Now, only 5 more to go!

Workflows?

Create a QC workflow that does the trimming

Or, cheat and import trimmed+paired datasets from the

RNA-Seq UCDavis 2013 Example Data →

Reads, Post-QC, Re-Paired

shared data library

NGS Data Quality: Further reading & Resources

FastQC Documenation

Read Quality Assessment & Improvement
by Joe Fass
From the UC Davis 2013 Bioinformatics Short Course

Manipulation of FASTQ data with Galaxy
by Blankenberg, *et al*.

Mapping with Tophat

RNA-Seq: Mapping with Tophat

Create new history

⚙ (cog) → Create New

Get filtered reads

Shared Data → Data Libraries

→ RNA-Seq UCDavis 2013 Example Data*

→ **Reads, Post-QC**

→ Select **MeOH_REP1_R1, MeOH_REP1_R2**

Also select **genes_chr12.gtf**

And then Import to current history

UCDAVIS Bioinformatics Core
Genome Center

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

RNA-seq Exercise: Mapping with Tophat

- 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

- Use Own Junctions → Yes
    - Use Gene Annotation → Yes
    - Gene Model Annotation → genes_chr12.gtf
- 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

TopHat Manual

Mapping with 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.

TopHat Manual

RNA-Seq Mapping With Tophat: Resources

RNA-Seq Concepts, Terminology, and Work Flows

by Monica Britton

Aligning PE RNA-Seq Reads to a Genome

by Monica Britton

both from the UC Davis 2013 Bioinformatics Short Course

RNA-Seq Analysis with Galaxy

by Jeroen F.J. Laros, Wibowo Arindrarto, Leon Mei

from the GCC2013 Training Day

RNA-Seq Analysis with Galaxy

by Curtis Hendrickson, David Crossman, Jeremy Goecks

from the GCC2012 Training Day

RNA-Seq: Differential Expression with Cuffdiff

RNA-Seq Differential Expression: Get the Data

Create new history

⚙ (cog) → Create New

Import:

Shared Data → Data Libraries

→ RNA-Seq UCDavis 2013 Example Data*

→ Tophat Outputs

→ Select all accepted_hits datasets

Also select **genes_chr12.gtf**

And then Import to current history

UCDAVIS Bioinformatics Core
Genome Center

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

Cuffdiff

- Part of the Tuxedo RNA-Seq Suite (as are Tophat and Bowtie)
- Identifies differential expression between multiple datasets
- Widely used and widely installed on Galaxy instances

NGS: RNA Analysis → Cuffdiff

Cuffdiff

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

Cuffdiff

- Running with 2 Groups: MeOH and R3G
- Each group has 3 replicates each

Cuffdiff

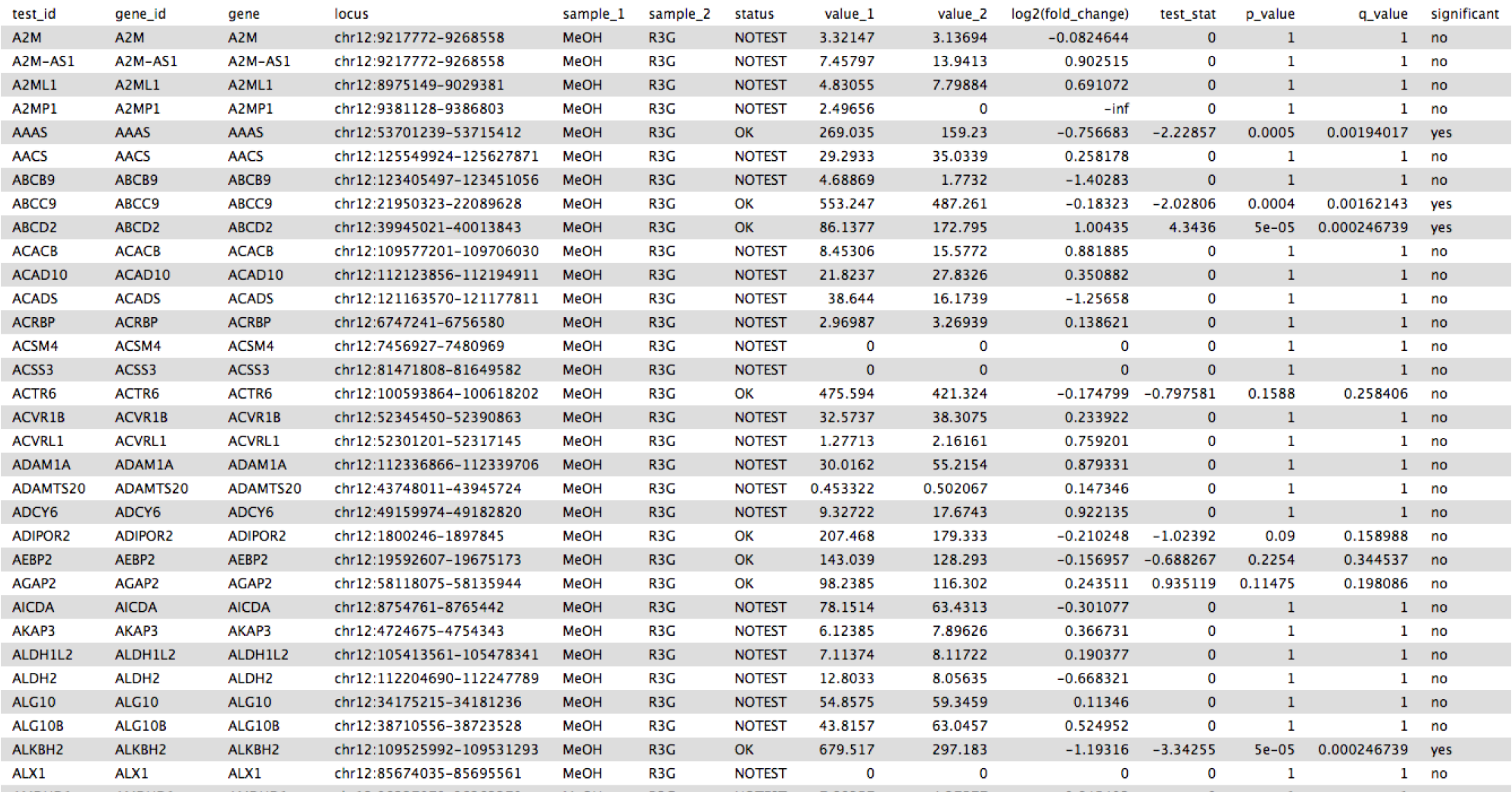
- Which Transcript definitions to use?
  - Official (genes_chr12.gtf in our case)
  - MeOH or R3G Cufflinks transcripts
  - Results of Cuffmerge on MeOH & R3G Cufflinks transcripts
- Depends on what you care about

NGS: RNA Analysis → Cuffdiff

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 in FPKM
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?

RNA-Seq Differential Expression with Cuffdiff: Resources

RNA-Seq Concepts, Terminology, and Work Flows

by Monica Britton

from the UC Davis 2013 Bioinformatics Short Course

RNA-Seq Analysis with Galaxy

by Jeroen F.J. Laros, Wibowo Arindrarto, Leon Mei

from the GCC2013 Training Day

RNA-Seq Analysis with Galaxy

by Curtis Hendrickson, David Crossman, Jeremy Goecks

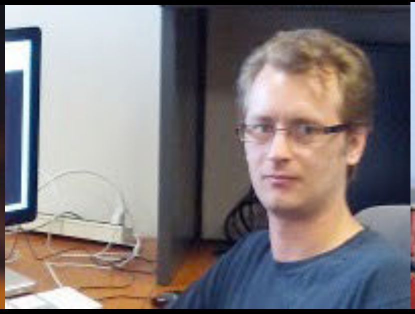
from the GCC2012 Training Day

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

Galaxy is hiring post-docs and software engineers

Please help.

http://wiki.galaxyproject.org/GalaxyIsHiring

Also Thanks To

VSU VIRGINIA STATE

College of Natural
and Health Sciences

Glenn Harris

National Institutes of Health
Amazon Web Services

Thanks

Dave Clements

Galaxy Project

Johns Hopkins University

outreach@galaxyproject.org