

# GO Galaxy

---

6th International Biocuration  
Conference  
8 April 2013, Cambridge, UK

**Suzanna Lewis, Chris Mungall, Seth Carbon**

Lawrence Berkeley National Lab

<http://www.berkeleybop.org/>

**Amelia Ireland, University of California Berkeley**

<http://gmod.org/>

**Dave Clements, Emory University**

<http://galaxyproject.org/>



# Agenda

15:00 Galaxy Overview &  
Basic Analysis with Galaxy

16:00 Ontological Analysis with GO Galaxy

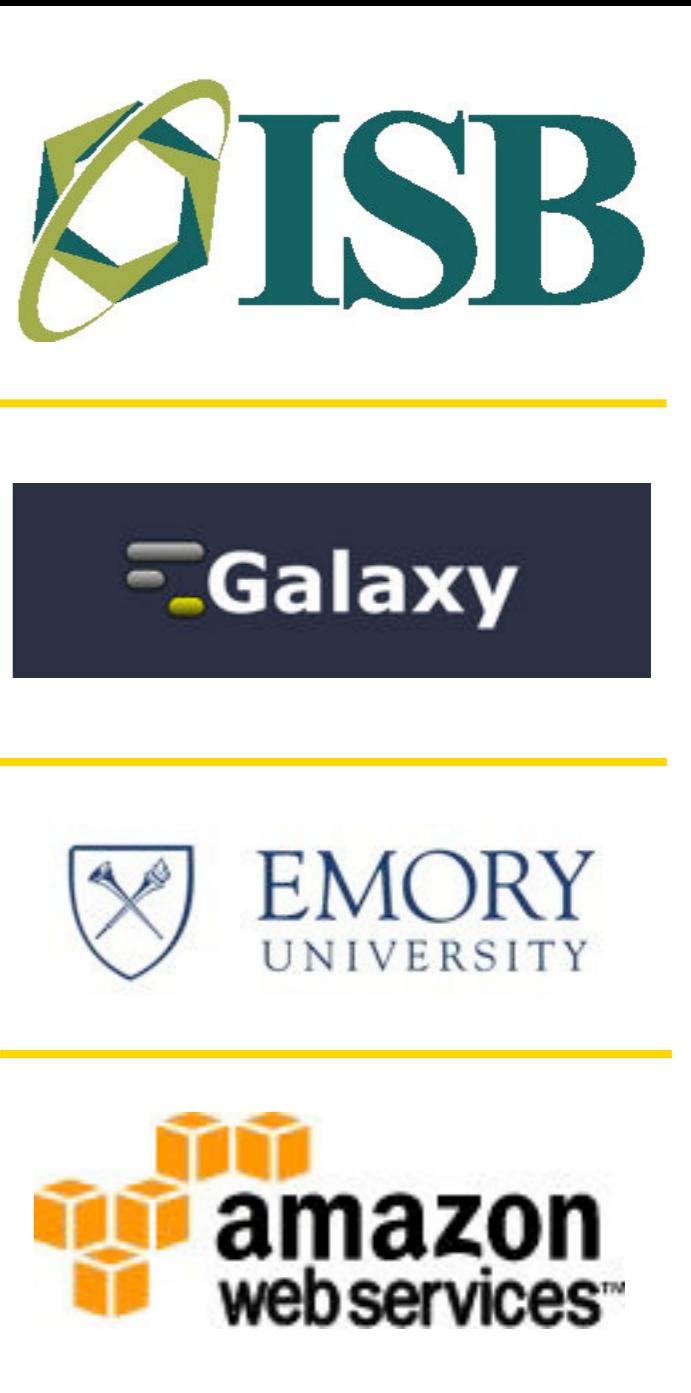
17:00 Done

# Introduction to Galaxy

---

6th International Biocuration  
Conference  
8 April 2013, Cambridge, UK

Dave Clements, Emory University  
<http://galaxyproject.org/>



# Agenda

Basic Analysis with Galaxy

Basic Analysis into Reusable Workflows

Sharing: How to Play Well with Others

Galaxy Project Overview

Further reading

# What is Galaxy?

# What is Galaxy?

Let's **answer that by demonstrating it**

- Please follow along
- Please **share the joy with your neighbors**. Not all of you have laptops
- We will have **wandering TA's** to help; just flag them down.
- Given size of the audience, and the time limits **the presenter won't wait for everyone to catch up**.

# What is Galaxy?

Let's **answer that by demonstrating it**

- Please follow along
- Please **share the joy with your neighbors**. Not all of you have laptops
- We will have **wandering TA's** to help; just flag them down.
- Given size of the audience, and the time limits **the presenter won't wait for everyone to catch up**.

**"you'll definitely want to have a plan B in case the network explodes"**

a Biocuration 2013 organizer

# Basic Analysis

On human chromosome 22,  
which coding exons have the most  
repeats in them?

cloud1.galaxyproject.org (I)  
cloud2.galaxyproject.org (S)  
cloud3.galaxyproject.org (B)

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

# Exons & Repeats: A General Plan

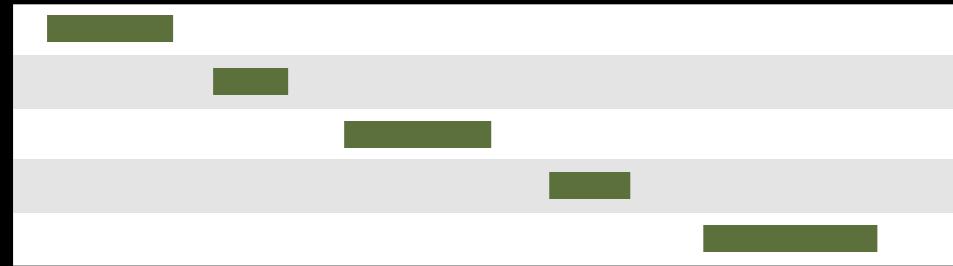
- Get some data
  - Coding exons on chromosome 22
  - Repeats on chromosome 22
- Mess with it
  - Identify which exons have repeats
  - Count repeats per exon
  - Save, download, ... exons with most repeats

cloud1.galaxyproject.org (I)

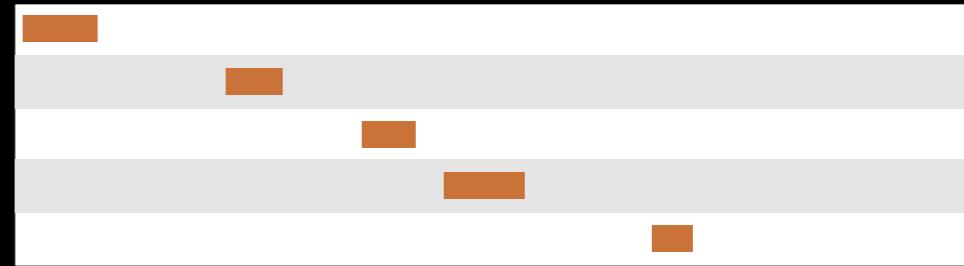
cloud2.galaxyproject.org (S)

cloud3.galaxyproject.org (B)

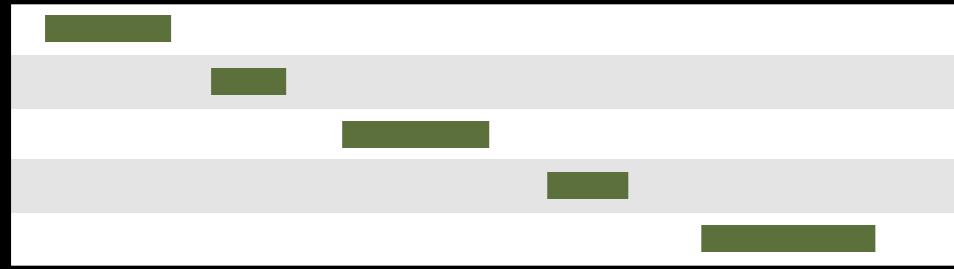
(~ <http://usegalaxy.org/galaxy101> )



Exons, from UCSC



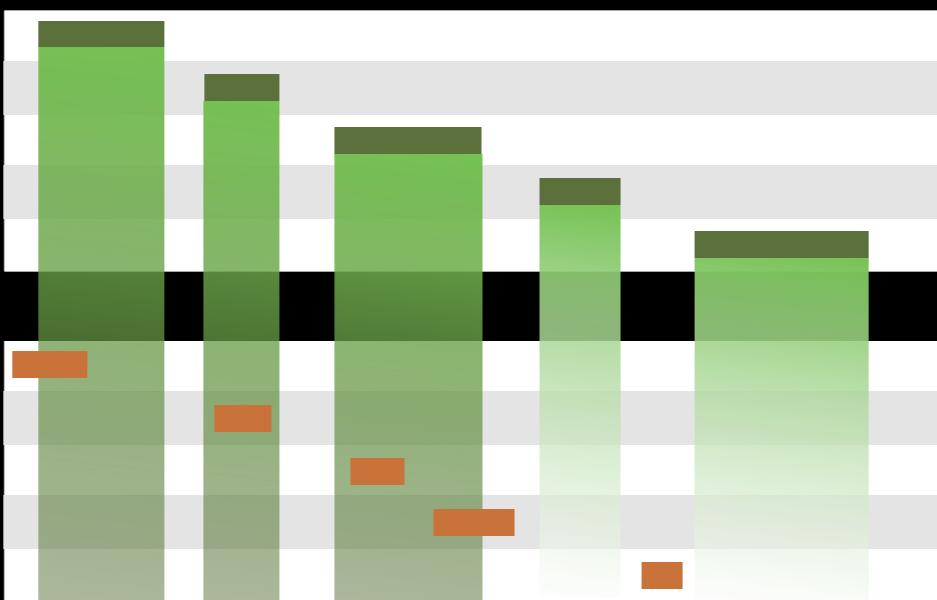
Repeats, from UCSC



Exons, from UCSC



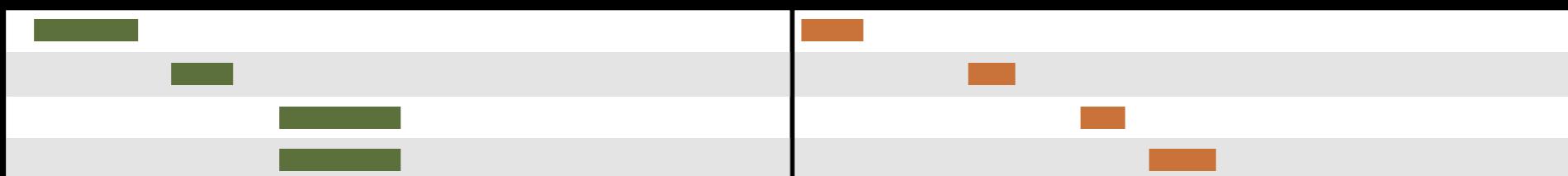
Repeats, from UCSC

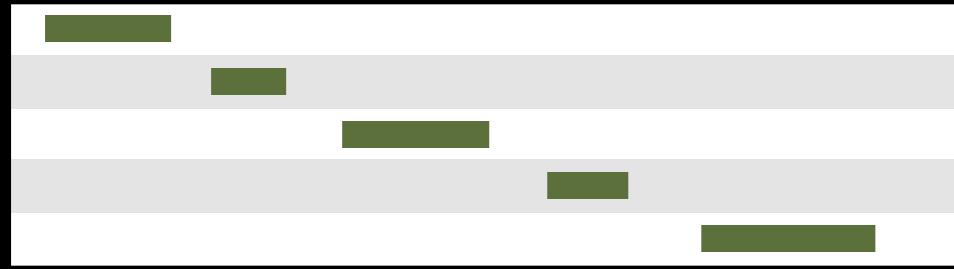


Exons, from UCSC

Repeats, from UCSC

Overlap pairings

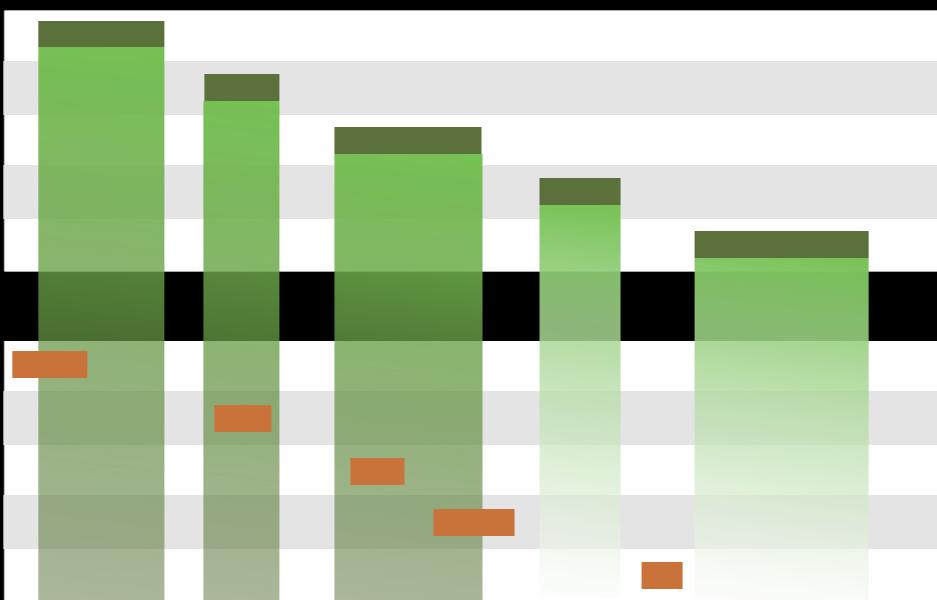




Exons, from UCSC



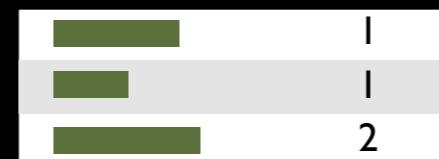
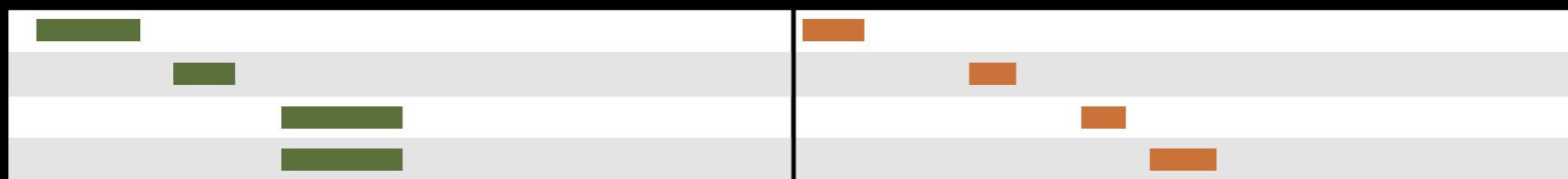
Repeats, from UCSC



Exons, from UCSC

Repeats, from UCSC

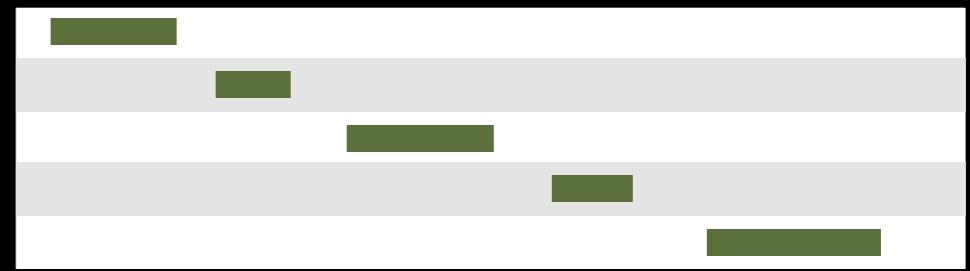
Overlap pairings



Exon overlap counts



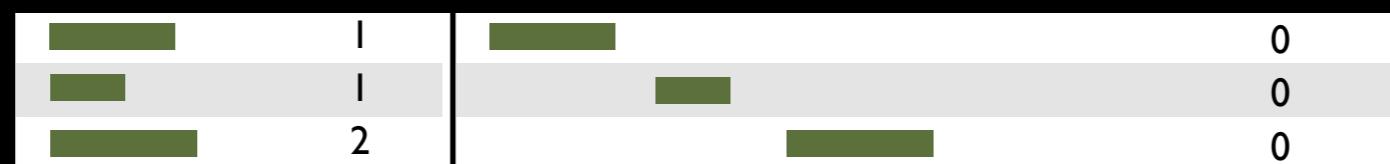
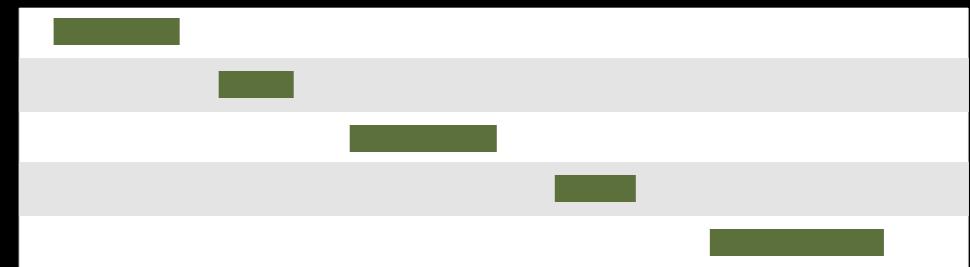
Exon overlap counts



Exons, from UCSC



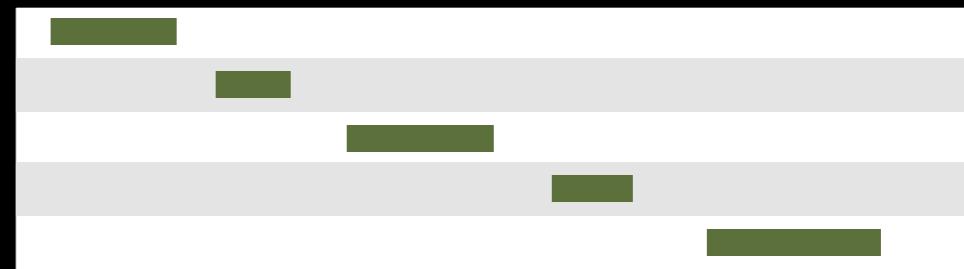
Exon overlap counts



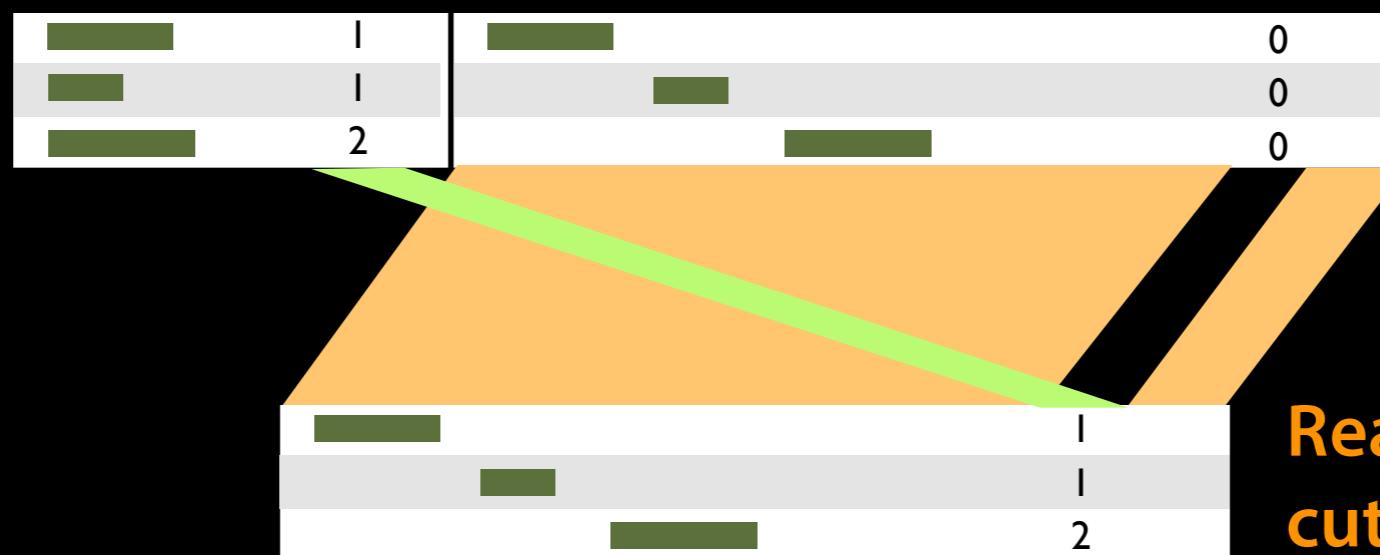
Join on exon name

█	1
█	1
█	2

Exon overlap counts



Exons, from UCSC



Join on exon name

Rearrange columns w/  
cut

# Agenda

Basic Analysis with Galaxy

**Basic Analysis into Reusable Workflows**

Sharing: How to Play Well with Others

Galaxy Project Overview

Further reading

# Exons and Repeats *History* → Reusable Workflow?

- The analysis we just finished was about
  - Human chromosome 22
  - Overlap between exons and repeats
- But, ...
  - there is **nothing inherently** in the analysis **about humans, chromosomes, 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 generic *Overlap* Workflow

## Extract Workflow from history

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

## Run / test it

*Tonight at a pub*

Count # CpG islands in each exon

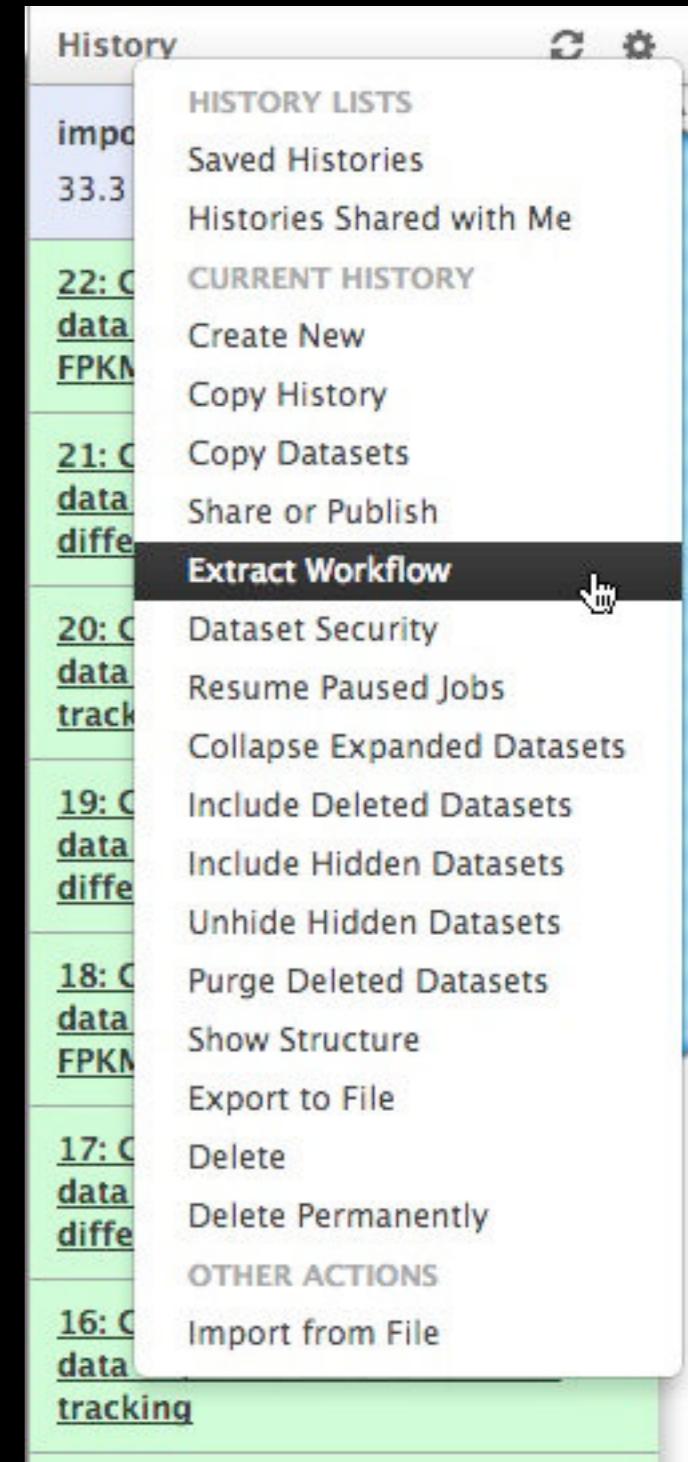
Did that work?

*And*

Count # of exons in each repeat

Did that work? *Why not?*

Edit workflow: doc assumptions



# Agenda

Basic Analysis with Galaxy

Basic Analysis into Reusable Workflows

**Sharing: How to Play Well with Others**

Galaxy Project Overview

Further reading

# Sharing & Publishing enables Reproducibility

~~We~~ather Reproducibility: Everybody talks about it, but ...

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

All analysis and metadata 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



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:    
[Advanced Search](#)

## Windshield splatter analysis with the Galaxy metagenomic pipeline

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

### OPEN ACCESS ARTICLE

#### This Article

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

Copyright © 2009 by Cold  
Spring Harbor Laboratory  
Press

- » Abstract [Free](#)
- » Full Text (PDF) [Free](#)

### Current Issue

October 2010, 20 (10)



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

# Sharing & Publishing enables Reproducibility



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](#)

## Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond<sup>1,2,6,9</sup>, Samir Wadhawan<sup>3,6,7</sup>,

Francesca Chiaromonte<sup>4</sup>, Guruprasad Ananda<sup>1,3</sup>, Wen-Yu Chung<sup>1,3,8</sup>,

James T

### Footnotes

#### OPEN ACCESS ARTICLE

##### This Article

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

Copyright © 2009 by Cold Spring Harbor Laboratory Press

#### Current Issue

October 2010, 20 (10)



[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>.]

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

# Sharing for Galaxy Administrators Too

Data Libraries

Make data easy to find

Genome Builds

Care about a particular subset of life?

Galaxy Tool Shed

Wrapping and sharing tools and datatypes

# Agenda

Basic Analysis with Galaxy

Basic Analysis into Reusable Workflows

Sharing: How to Play Well with Others

Galaxy Project Overview

Further reading

# What is Galaxy, and how can I get some?

- An open, web-based platform for **accessible**, **reproducible**, and **transparent** computational biomedical research.
- Galaxy is available as:
  - **A free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
  - **Open source software** that makes integrating your own tools and data and customizing for your own site simple
  - **Free cloud images** that can be deployed by informatics novices

<http://galaxyproject.org>

# A free for everyone web-based service: [usegalaxy.org](http://usegalaxy.org)

Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User

Using 3%

Tools

search tools

[Get Data](#)

[Send Data](#)

[ENCODE Tools](#)

[Lift-Over](#)

[Text Manipulation](#)

[Convert Formats](#)

[FASTA manipulation](#)

[Filter and Sort](#)

[Join, Subtract and Group](#)

[Extract Features](#)

[Fetch Sequences](#)

[Fetch Alignments](#)

[Get Genomic Scores](#)

[Operate on Genomic Intervals](#)

[Statistics](#)

[Graph/Display Data](#)

[Regional Variation](#)

[Multiple regression](#)

[Multivariate Analysis](#)

[Evolution](#)

[Motif Tools](#)

[Multiple Alignments](#)

[Metagenomic analyses](#)

[Genome Diversity](#)

[Phenotype Association](#)

[EMBOSS](#)

NGS TOOLBOX BETA

[NGS: QC and manipulation](#)

[NGS: Mapping](#)

[NGS: SAM Tools](#)

[NGS: GATK Tools \(beta\)](#)

[NGS: Variant Detection](#)

NIOO  
nbIC SURF SARA BiG Grid the dutch e-science grid  
Andromeda: A cloud-based Galaxy

Live Quickies

Basic fastQ manipulation: Galactic quickie # 13

Advanced fastQ manipulation: Galactic quickie # 14

454 Mapping: Single End Galactic quickie # 15

Uploading Data using FTP Galactic quickie # 17

Managing account histories Galactic quickie # 19

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 CyberScience at Penn State](#), and [Emory University](#).

Galaxy build: \$Rev: 8778:7c3df0bcbe22\$

galaxyproject

intermineorg Take a look at our new interactive web services docs: [docs.labs.intermine.org/flymine-beta](http://docs.labs.intermine.org/flymine-beta)  
15 hours ago · reply · retweet · favorite

galaxyproject Jackson Lab surveying bioinformatics cores. Scientific computing [svy.mk/X905zC](http://svy.mk/X905zC) Bioinformatics and stats [svy.mk/W7637u](http://svy.mk/W7637u)  
15 hours ago · reply · retweet · favorite

galaxyproject GCC2013 registration and abstract submission are now open [bit.ly/GCC2013Reg](http://bit.ly/GCC2013Reg) [bit.ly/gcc2013abs](http://bit.ly/gcc2013abs) #usegalaxy  
yesterday · reply · retweet · favorite

more ...

This is a free, public, [internet accessible resource](#). Data transfer and data storage are not encrypted. If there are restrictions on the way your research data can be stored and used, please consult your local institutional review board or the project PI before uploading it to any public site.

History

Full dataset for CPB Chip-Seq protocol  
6.9 GB

10: FastQC Filter FASTQ on data 7.html

9: Filter FASTQ on data 7

8: FastQC FASTQ Groomer on data 5.html

7: FASTQ Groomer on data 5

6: FASTQ Groomer on data 5

5: Mouse ChIP-Seq Example Experimental Data, chr19\_mm9

4: Mouse ChIP-Seq example Control Data, chr19\_mm9

3: FastQC FASTQ Groomer on data 1.html

2: FASTQ Groomer on data 1

1: http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeSydhTfbs/wgEncodeSydhTfbsMeCtcfDmso20lggyaleRawDataRep1.fastq

(See also: <http://bit.ly/gxyServers>)

However, *a centralized solution cannot scale to meet the analysis needs of the entire world.*

# Open Source Software: [getgalaxy.org](http://getgalaxy.org)

- Galaxy is designed for local installation and customization
  - Easily integrate new tools
  - Requires a computational resource on which to be deployed

<http://getgalaxy.org>

# Galaxy is available *on the cloud*

- Using this today
- 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
- **We are using this today**



<http://usegalaxy.org/cloud>

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

# Galaxy Resources and Community: Mailing Lists

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

## Galaxy-Announce

Project announcements, low volume, moderated

Low volume ( 42 posts, 1600 members in 2012)

## Galaxy-User

Questions about using Galaxy and usegalaxy.org

High volume (2900 posts, 2700 members in 2012)

## Galaxy-Dev

Questions about developing for and deploying Galaxy

High volume (4500 posts, 850 members in 2012)

# Community can create, vote and comment on **issues**

The screenshot shows a Trello board titled "Galaxy: Development Inbox". The board is organized into four main columns:

- Inbox:** Contains cards for adding cards, filtering, sorting, uploaded fastq files, reference genome requests, feature requests, and an "Add a card..." button.
- Developer ideas:** Contains cards for anonymous workflow use, feature requests like restarting failed workflows, Google Drive/Dropbox integration, bug reports for tool tips and FASTQ summary statistics, standalone web applications, archiving histories, modifying data library upload completion messages, and displaying in UI runtime.
- Bug Reports:** Contains cards for issues with workflow step hiding, broken workflow views, unable to run jobs with job limits, bugs with data\_columns, Velvet wrapper problems, apport.fileutils, and functional tests for migrated or instantiated tools.
- Issues from Bitbucket:** Contains cards for enabling history creation, requiring history names, flexible output handlers, overriding parameters, a suggestion for a new tag in XML files, a real DB key build ontology, and adding password secure tools.

On the right side of the board, there are additional sections:

- Members:** Shows a grid of member profiles and an "Add Members..." button.
- Board:** Includes options for "Options", "Add List", and "Filter Cards".
- Activity:** Displays recent activity feed with entries from Dannon Baker, g2roboto, and others.

<http://bit.ly/gxytrello>



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 service web site 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 on [this wiki](#) and elsewhere.

[usegalaxy.org](http://usegalaxy.org)

### Deploy Galaxy

Galaxy is open source for all organizations. Local Galaxy servers can be set up by [downloading and customizing](#) the Galaxy application.

- Admin
- Cloud

[getgalaxy.org](http://getgalaxy.org)

### Community & Project

Galaxy has a large and active user community and many ways to [Get Involved](#).

- [Community](#)
- [News](#)
- [Events](#)
- [Support](#)
- [Galaxy Project](#)

### Contribute

- **Users:** Share your histories, workflows, visualizations, data libraries, and [Galaxy Pages](#), enabling others to use and learn from them.
- **Deployers and Developers:** Contribute tool definitions to the [Galaxy Tool Shed](#) (making it easy for others to use those tools on their installations), and code to the core release.
- **Everyone:** [Get Involved!](#)



[Talk abstracts](#)  
due **12 April**  
Be published in  
*GigaScience*

### Use Galaxy

[Use Main \(about\)](#)  
[Use Others!](#) • [Learn](#)  
[Share](#) • [Search](#)

### Communication

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

### Deploy Galaxy

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

### Contribute

[Tool Shed](#) • [Share](#)  
[Issues & Requests](#)  
[Support](#)

### Galaxy Project

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

# Events

# News

## Galaxy Wiki

Events

DaveClements Settings Logout | Search:

### 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 relevant to the Galaxy Community. This is also available as an RSS feed [RSS](#).

If you know of any event that should be added to this page and/or to the Galaxy Event Calendar, please add it here or send it to [outreach@galaxyproject.org](mailto:outreach@galaxyproject.org).

### Upcoming Events



Date	Topic/Event	Venue/Location
April 7-10	GO Galaxy Workshop	Biocuration 2013, Cambridge, United Kingdom
April 7-8	BOSC/Broad Interoperability Hackathon	Cambridge, Massachusetts, United States
April 9-11	Workshop: <i>Integrated Research Data Management for Next Gen Sequencing Analysis Using Galaxy and Globus Online Software-as-a-Service</i>  Talk: <i>Integrated Research Data management and Analysis in NGS using Globus Online, Galaxy and Amazon Web Services</i>	BioIT World, Boston, Massachusetts, United States
April 10	Introduction to Galaxy Boot Camp  Registration is full	UC Davis Bioinformatics Core Davis, California, United States
April 11	Introduction to RNASeq Boot Camp  Registration is full	UC Davis Bioinformatics Core Davis, California, United States
April 11	Introduction to Galaxy Workshop	The Genome Analysis Centre, Norwich, United Kingdom
April 12	Next generation sequencing data interpretation: enhancing reproducibility and accessibility	Reed College, Portland, Oregon, United States
May 14-16	Tutorial: Exploring and Enabling Biomedical Data Analysis with Galaxy	Great Lakes Bioinformatics Conference (GLBIO) 2013, Pittsburgh, Pennsylvania, United States
May 16-17	Galaxy Workflows for Bioinformatics Analysis, and Workshop 1A – Galaxy Workflows for Bioinformatics Analysis	Workshop in Next-Generation Sequence Analysis and Metabolomics (WiNGS), UNC-Charlotte, North Carolina, United States
May 21 May 29	Initiation à l'utilisation de Galaxy  Les deux ateliers sont maintenant complets	Cycle "Bioinformatique par la pratique" 2013, INRA Jouy-en-Josas, France
May 22	Analyse de données issues de séquenceurs nouvelle génération sous Galaxy	

## News

Announcements of interest to the Galaxy Community. These can include items from the Galaxy Team or the Galaxy community and can address anything that is of wide interest to the community.

The Galaxy News is also available as an RSS feed [RSS](#).

See [Add a News Item](#) below for how to get an item on this page, and the RSS feed. Older news items are available in the [Galaxy News Archive](#).

### See also

- [Distribution News Briefs](#)
- [Galaxy Updates](#)
- [Galaxy on Twitter](#)
- [Events](#)
- [Learn](#)
- [Support](#)
- [About the Galaxy Project](#)

## News Items

- February 2013 Galaxy Update
- GCC2013 Training Day Topics: Vote!
- Galaxy Project Openings
- Jan 11, 2013 Distribution & News Brief
- January 2013 GalaxyAdmins
- January 2013 Galaxy Update
- Dec 20, 2012 Distribution & News Brief
- Galaxy Internships @ EMBL
- Nominate GCC2013 Training Topics
- Dec 3, 2012 Distribution & News Brief
- December 2012 Galaxy Update
- Nov 14, 2012 Distribution & News Brief
- NGS Analysis by Viz. with Trackster
- November 2012 GalaxyAdmins

[News Archive](#)

## News Items

### February 2013 Galaxy Update

The February 2013 Galaxy Update is now available.

#### Highlights:

- Three new public Galaxy servers
- New papers
- Open Positions at five different institutions
- GCC2013 Training Day Topic voting, Registration, and Sponsorships
- January GalaxyAdmins Web Meetup slides and screencast
- Other Upcoming Events and Deadlines
- Galaxy Distributions
- Tool Shed Contributions
- Other News

If you have anything you would like to see in the March Galaxy Update, please let us know.

Dave Clements and the Galaxy Team

*Posted to the Galaxy News on 2013-02-01*

### GCC2013 Training Day Topics: Vote!

A list of possible topics for the GCC2013 Training Day is now available. Please take a few minutes to review these possibilities and then vote for your favorite three topics.\*

Your votes will determine not only the topics that are offered, but also which topics should be offered more than once, assigned to which rooms, and which ones should not be scheduled at the same time. Your vote matters.



galaxyproject.org/GCC2013



STARTING  
@  
€95



galaxyproject.org/GCC2013



STARTING

@

€95

Talk abstracts due **12 April**

(GIGA)<sup>n</sup> SCIENCE



# Agenda

Basic Analysis with Galaxy

Basic Analysis into Reusable Workflows

Sharing: How to Play Well with Others

Galaxy Project Overview

Further reading

# Further Reading

Learn Galaxy Hub Page

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

Worked RNA-Seq Example

<http://bit.ly/GxyRNASEqEx>

Public Galaxy Server List

<http://bit.ly/gxyServers>

Community Wiki

<http://wiki.galaxyproject.org>

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

The screenshot shows the Galaxy Web Search interface. At the top, there is a header bar with the "Galaxy Web Search" logo and a search bar containing "Google™ Custom Search". To the right of the search bar are "Search" and "X" buttons. Below the header, a message says "Search the entire set of Galaxy web sites and mailing lists using Google." followed by a link "Run this search at Google.com (useful for bookmarking)". There is also a link "Want a different search?". At the bottom of the page is a link "Project home".

The screenshot shows the Galaxy Web Search results page for the query "chip-seq". The search bar at the top contains "chip-seq". Below the search bar is a navigation menu with tabs: All, Tools, Email, Source code, Shared, Documentation, Abstracts, and Requests. The "All" tab is selected. A message "About 444 results (0.06 seconds)" is displayed above the search results. The results themselves are listed below, each with an orange arrow pointing from its corresponding text label on the left to the result itself.

*Find*

Everything on ...

Tools for ...

Email about ...

Source code for ...

Published Histories, Pages, Workflows, about ...

Related feature requests

Papers using Galaxy for ...

Documentation on ...

# Acknowledgements

Francis Ouellette

Alex Bateman

Claire O'Donovan

Amy Cottage

Jennifer Harrow

Biocuration Organisers

Suzanna Lewis

Chris Mungall

Seth Carbon

Amelia Ireland

AWS Education Grant

NIH NSF Huck Institute

Penn State University Emory University

# The Galaxy Team



Enis Afgan



Dannon Baker



Dan Blankenberg



Dave Bouvier



Dave Clements



Nate Coraor



Carl Eberhard



Dorine Francheteau



Jeremy Goecks



Sam Guerler



Jen Jackson



Greg von Kuster



Ross Lazarus



Anton Nekrutenko



James Taylor

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

# Feedback Please!

**<http://bit.ly/GOISB2013>**

# Thanks



Dave Clements

Galaxy Project  
Emory University

[clements@galaxyproject.org](mailto:clements@galaxyproject.org)

<http://bit.ly/GOISB2013>