

Intro to Galaxy and the Galaxy Ecosystem

2013 Galaxy Community Conference

Training Day
June 30th, 2013



Anton Nekrutenko
Jennifer Hillman-Jackson
Penn State University



 **usegalaxy.org**

Goals

1. Introduce Galaxy
2. Introduce bioinformatics concepts and formats
3. Hands-on experience using a Cloud Galaxy
 - Load and integrate data
 - Perform bioinformatic analysis with Galaxy
 - Save, share, describe, and publish your analyses
 - Create, edit, and run a workflow
 - Visualize your results

Want more? Later see <http://usegalaxy.org>

Shared Data: Published Pages:

-->> **Many publications w/ tutorials**

-->> **Screencasts usegalaxy.org**

And see <http://galaxyproject.org> -->> **Learn**

Galaxy Project Mission

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.

Galaxy Project **Mission**

"Next-generation sequencing data interpretation: enhancing reproducibility and accessibility", by Nekrutenko & Taylor, *Nature Reviews Genetics*, 13, 667-672 (September 2012)

Galaxy as a Genomics WorkBench

Dataset:

Any input, output or intermediate set of data + metadata.
A record of a specific data or analysis step.

History:

A series of inputs, analysis steps, intermediate datasets, and outputs. A record of a group of data and analysis steps.

Tool:

An operation within Galaxy that acts upon dataset(s) as an analysis step. May be developed by Galaxy team or a 3rd party program that has been “wrapped” for Galaxy.

Workflow:

A series of analysis steps executed in a sequential stream.

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 and link to any Galaxy object (histories, datasets, workflows, visualization) or external resource (video, graphics, publications).

The tutorial we will do today is in a **Galaxy Page**.

Basic Analysis

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

Example has two key data manipulations:

- 1 - *coordinate join*: join based on overlapping genomic intervals
- 2 - *relational join*: join based on common keys between datasets

Plus other useful to know tasks:

importing histories, text manipulations, workflows, sharing

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

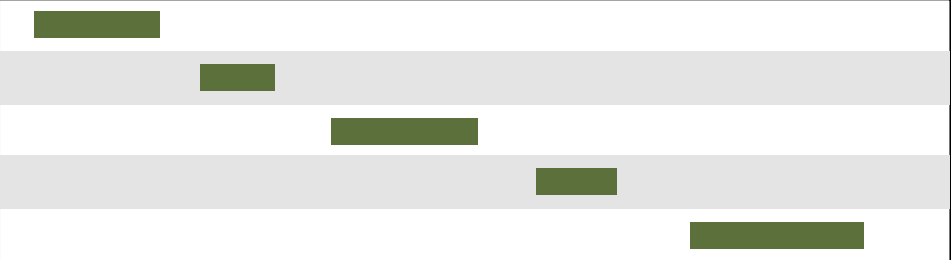
Exons & Repeats: A General Plan

- Get some data
 - Coding exons on chromosome 22
 - Repeats on chromosome 22
- Mess with it
 - Identify exons with repeats, count, rearrange data
 - Share, create/run workflow
 - Visualize Trackster & UCSC

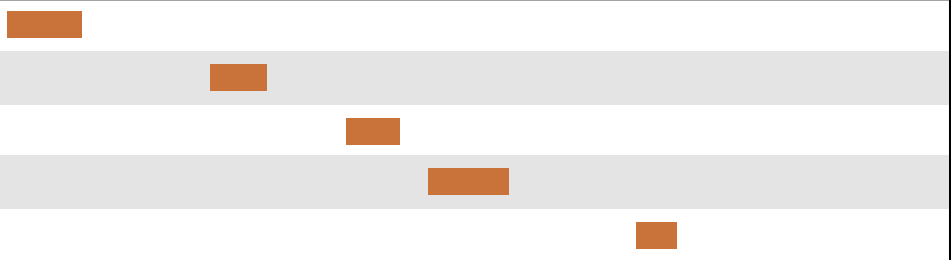
<http://cloud1.galaxyproject.org/>

<http://cloud2.galaxyproject.org/>

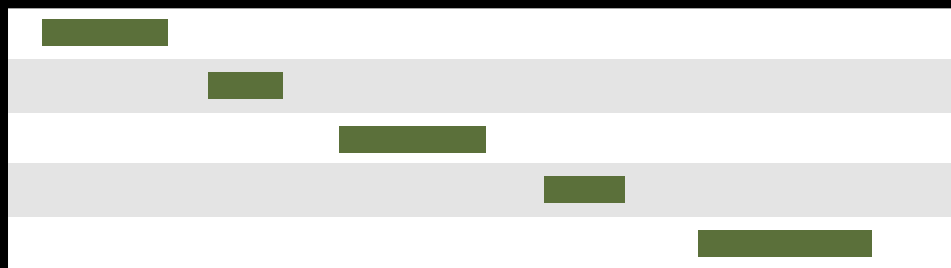
~ <http://usegalaxy.org/galaxy101>



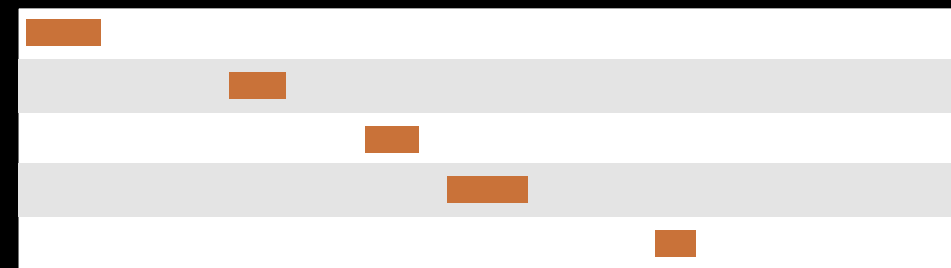
Exons, from UCSC



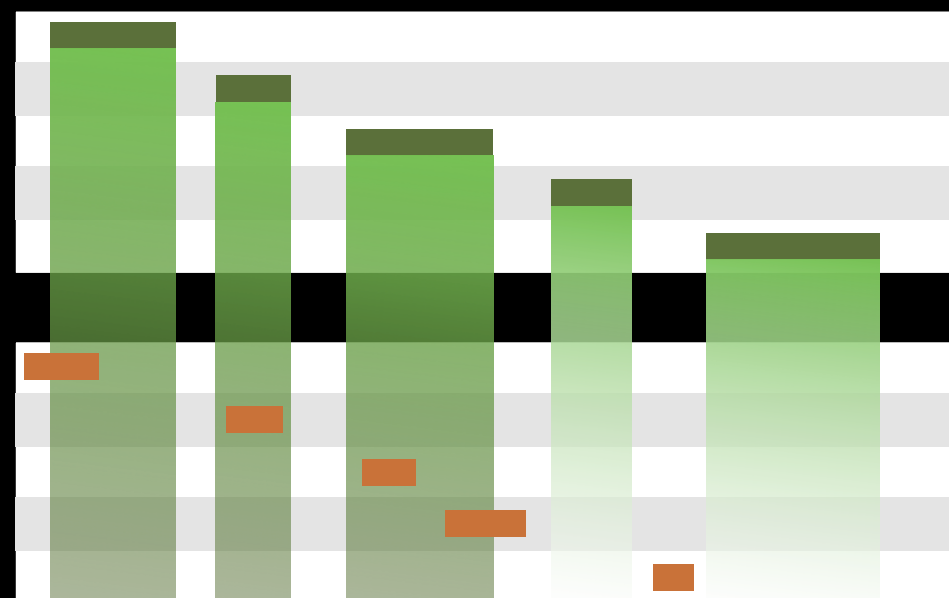
Repeats, from UCSC



Exons, from UCSC



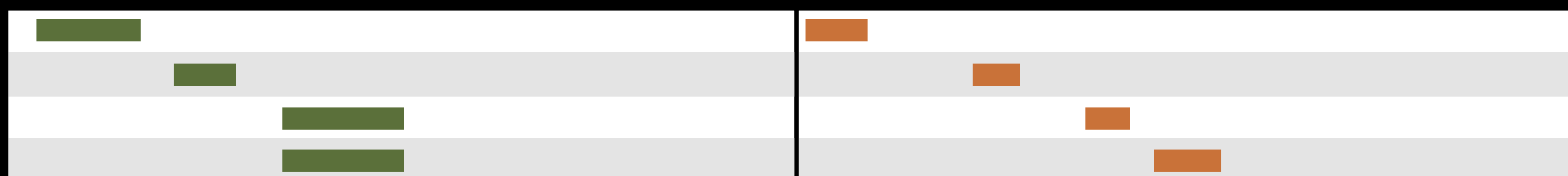
Repeats, from UCSC

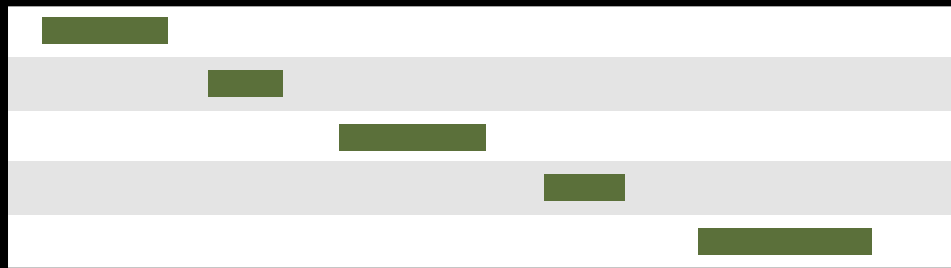


Exons, from UCSC

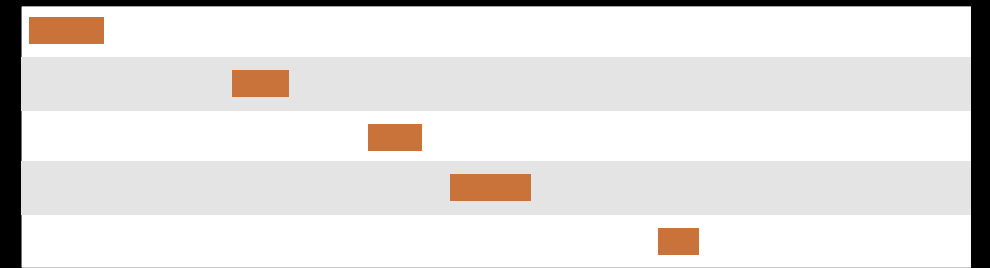
Repeats, from UCSC

Overlap pairings

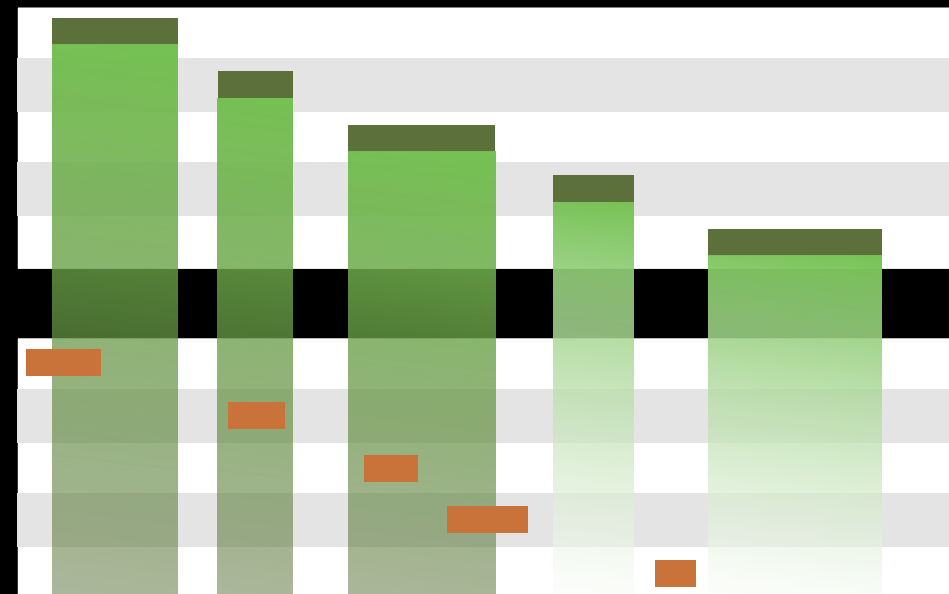




Exons, from UCSC



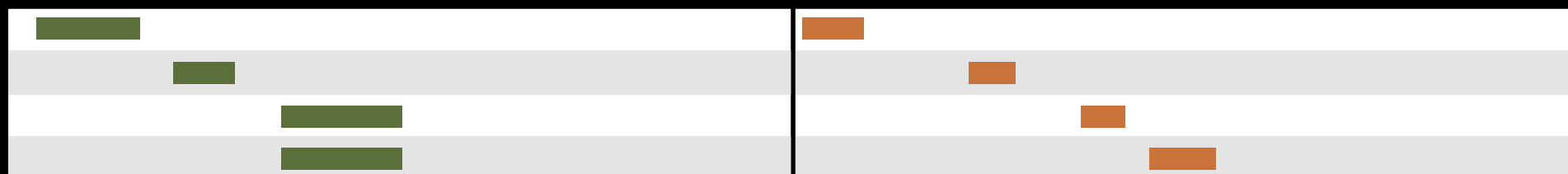
Repeats, from UCSC






Exons, from UCSC

Repeats, from UCSC

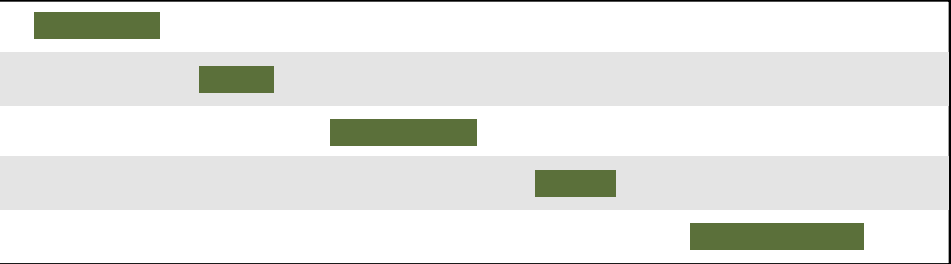
Overlap pairings




Exon overlap counts

	1
	1
	2

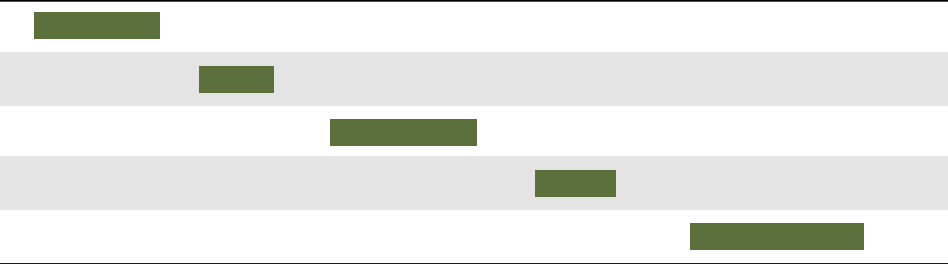
Exon overlap counts



Exons, from UCSC

	1
	1
	2

Exon overlap counts








Exons, from UCSC

	1		0
	1		0
	2		0




Join on exon name

	1
	1
	2

Exon overlap counts

Exons, from UCSC

	1		0
	1		0
	2		0

Join on exon name

	1
	1
	2

Rearrange columns w/
cut

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

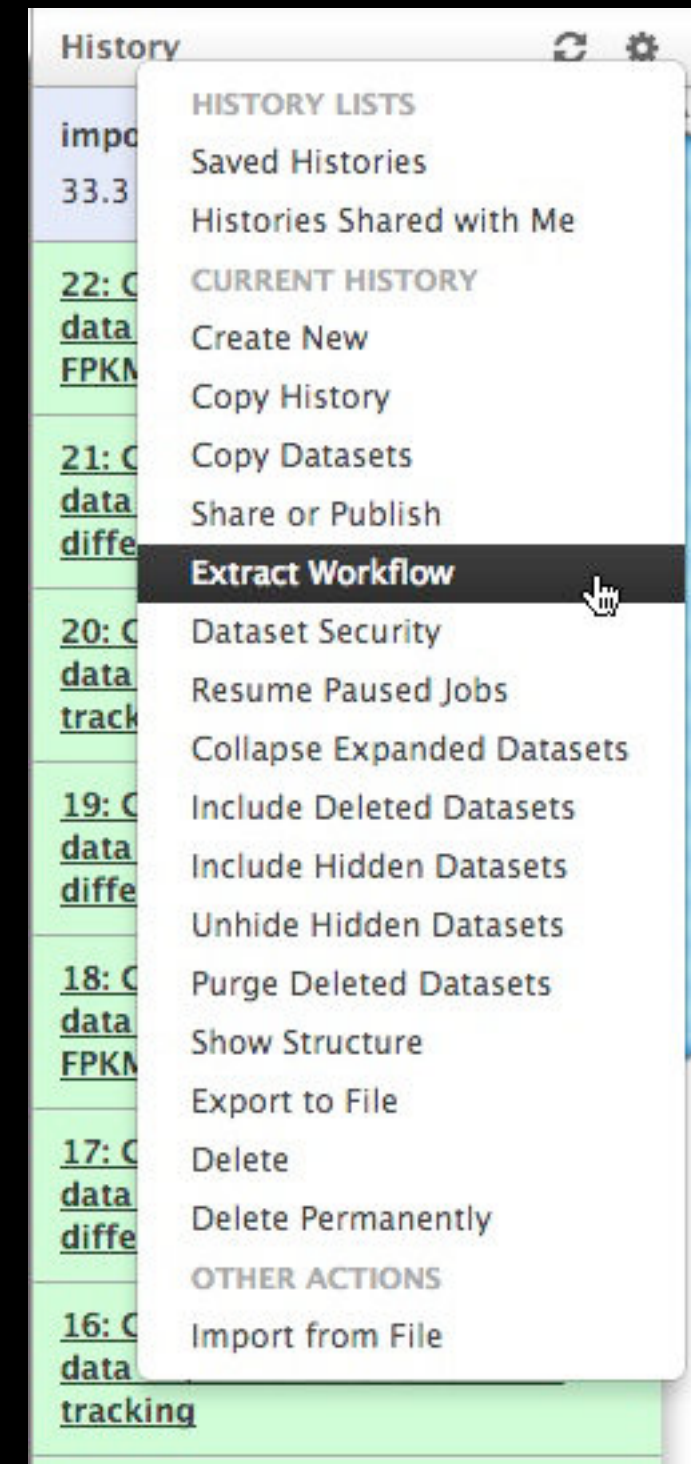
Guided: rerun with same inputs

On your own:

Use workflow with other inputs.
Count #SNPs in each coding
exon (see 101 tutorial).

On your own:

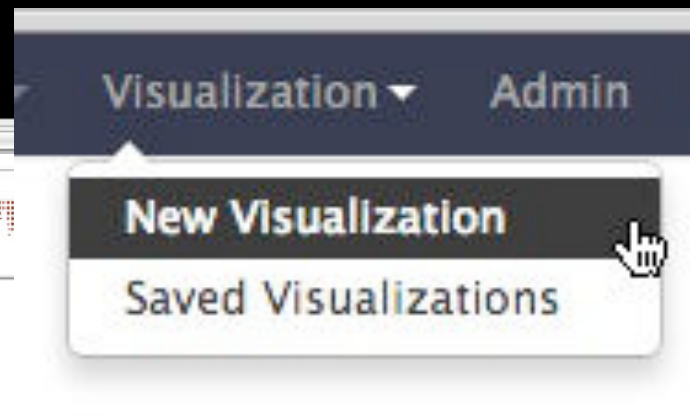
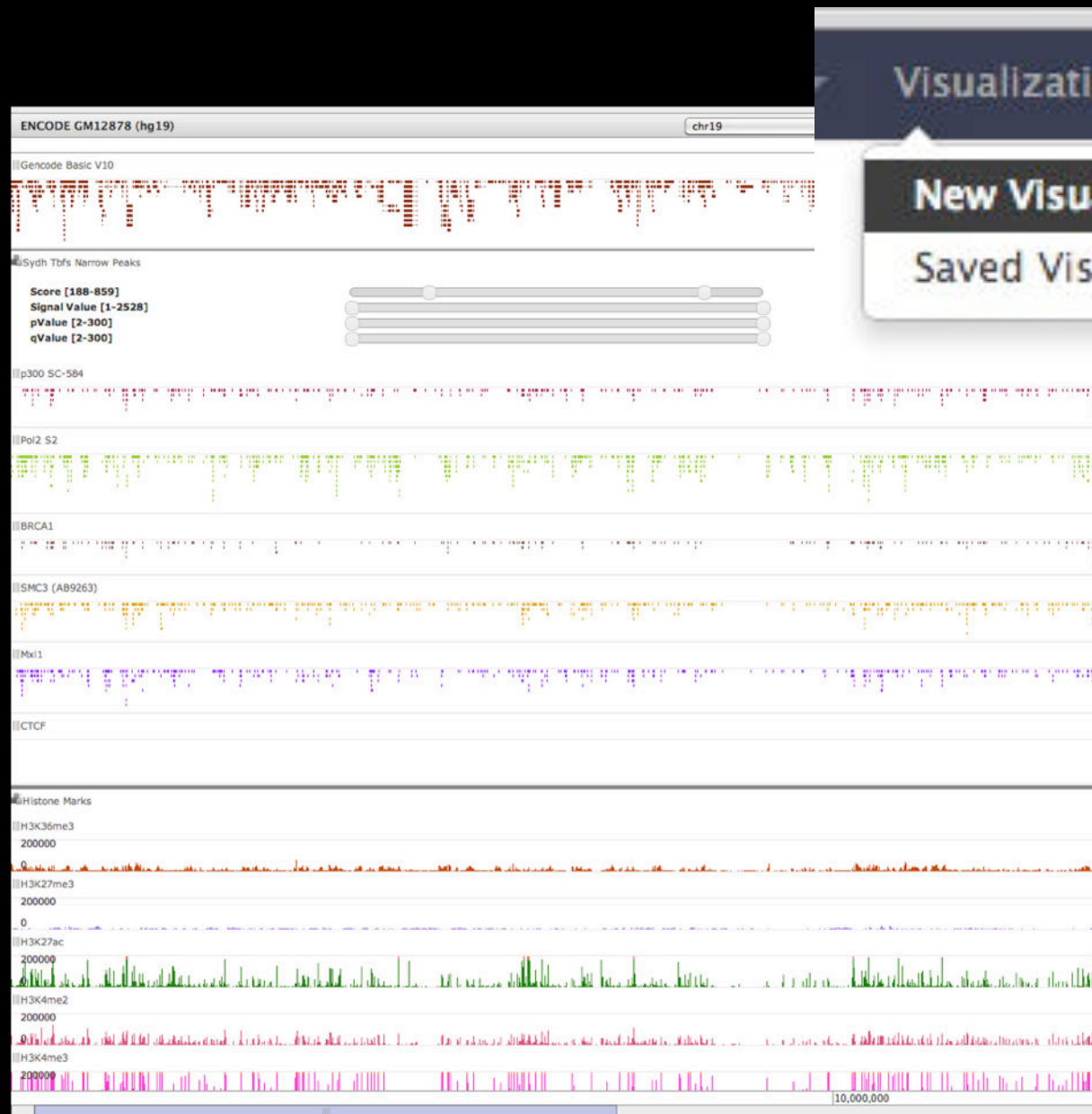
Can you add up transcripts with
most repeats? Chromosomes?



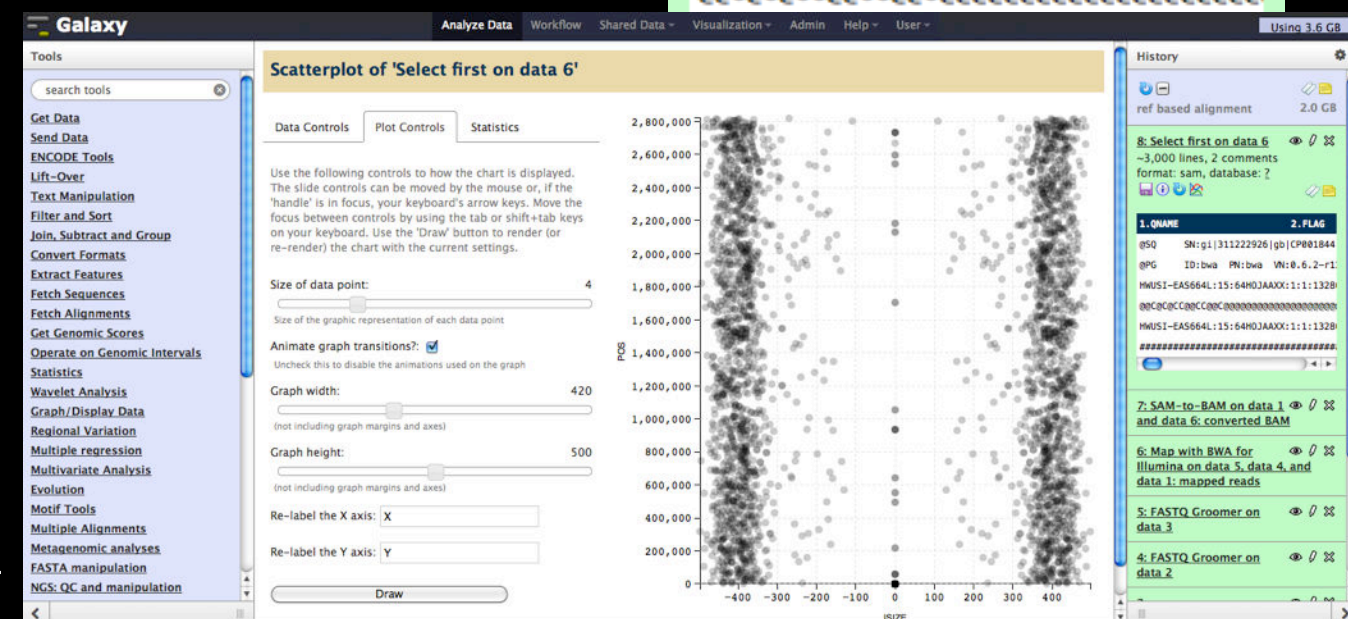
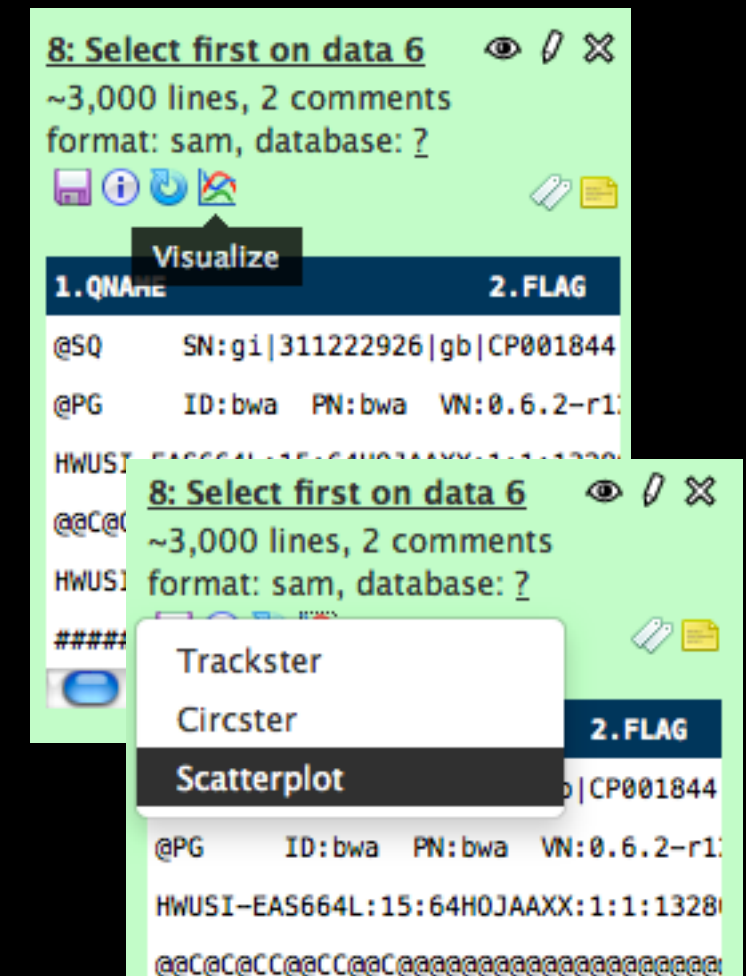
Exons and Repeats *History* → Reusable *Workflow*?

- 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 **determine the number of overlapping features of one dataset versus another, then assigns that value as the “score”, ending with a dataset otherwise in the same format and content as the original.**

Create a visualization in Galaxy



or



Jeremy Goecks, Nate Coraor, The Galaxy Team, Anton Nekrutenko & James Taylor, "NGS analyses by visualization with Trackster."
Nature Biotechnology 30, 1036–1039 (2012)

Vizualization inside Galaxy

- Leverages visualization to **evaluate and refine analyses**
- **Exposes basic analyses in visualization** to make it more informative
- Makea that **analyze-visualize-refine** loop seamless and **fast**
- Enables **learning tools and exploring their parameter space**
- Supports custom genome browsers, **without a predefined reference genome**

What is Galaxy?

- **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
- These options result in several **ways to use Galaxy**

<http://galaxyproject.org>

Using Galaxy - 4 ways

- **Public Main** Galaxy web instance: *usegalaxy.org*
- **Local** instance: *getgalaxy.org*
- **Cloud** instance: *usegalaxy.org/cloud*
- **Other Public** Galaxy web instances hosted by various groups:
wiki.galaxyproject.org/PublicGalaxyServers

<http://wiki.galaxyproject.org/Big%20Picture/Choices>

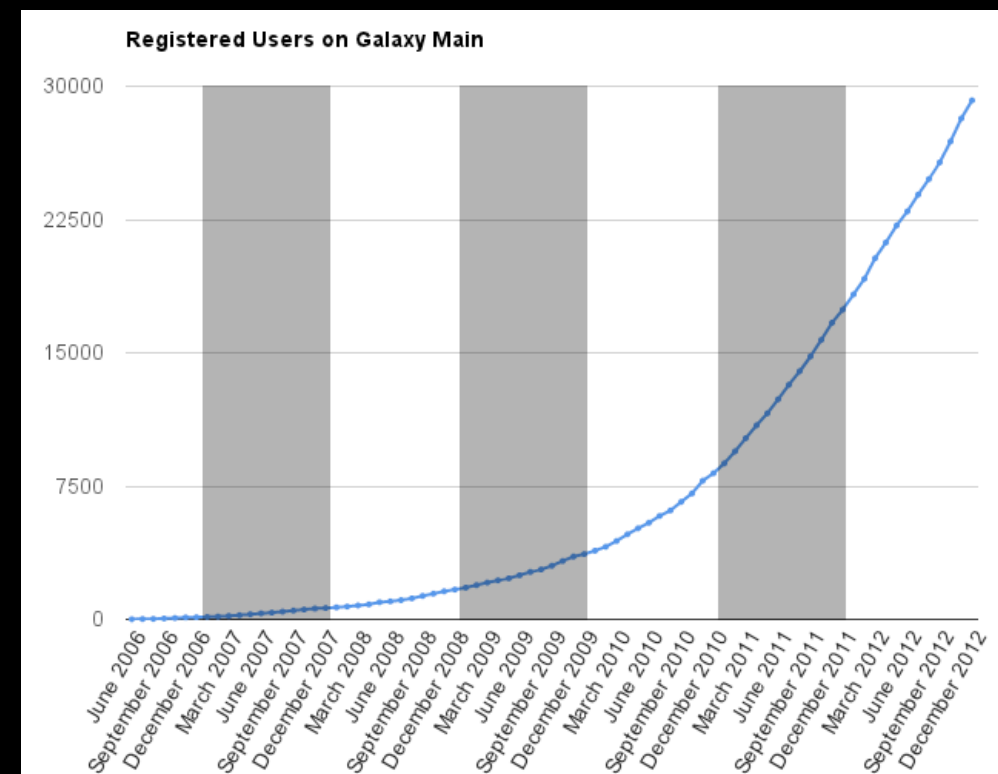
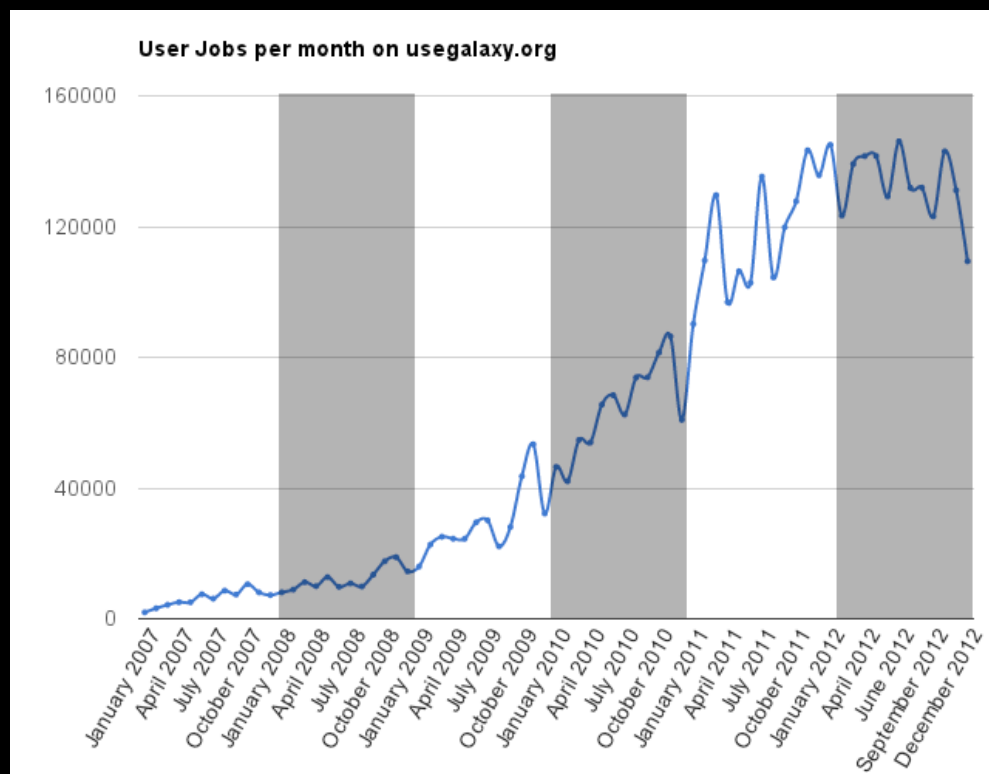
Galaxy is available ...

As a free (for everyone) web service

Galaxy “Main”

<http://usegalaxy.org>

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



Galaxy is available ...

- As a free (for everyone) web service

<http://usegalaxy.org>

- As open source software

<http://getgalaxy.org>

Galaxy “Local”

wiki.galaxyproject.org/DevNewsBriefs

galaxy-dist.readthedocs.org

bitbucket.org/galaxy/galaxy-dist

As Open Source Software: Local Galaxy Instances

- Galaxy is designed for local installation and customization
 - Easily integrate new tools
 - Easy to deploy and manage on nearly any (unix) system
 - Run jobs on existing compute clusters
- Requires a computational resource on which to be deployed

<http://getgalaxy.org>

Encourage **Local** Galaxy Instances

- Encourage and support Local Galaxy Instances
- Support **increasingly decentralized model** and improve access to existing resources
- Focus on building **infrastructure to enable the community to integrate and share** tools, workflows, and best practices

Galaxy Tool Shed

<http://toolshed.g2.bx.psu.edu>

The screenshot shows the Galaxy Tool Shed interface for the 'clustalomega' repository. The left sidebar contains links for 'Repositories', 'Browse by category', 'Browse all repositories', and 'Login to create a repository'. The main content area displays the 'Repository revision' section with a dropdown menu showing '2:bb1847435ec1'. Below this, the 'clustalomega' repository details are shown, including the clone URL, name, synopsis, detailed description, revision, owner, and times downloaded. A table at the bottom lists tools available in the repository.

name	description	version	requirements
Clustal Omega	multiple sequence alignment program for proteins	1.0.2	none

The screenshot shows the Galaxy Tool Shed interface displaying a list of repositories. The left sidebar is the same as in the previous screenshot. The main content area shows a search bar and a table of repositories.

Name	Synopsis	Revision	Category	Owner
abyss_toolsuite	This suite contains Abyss and Abyss-PE config files and wrappers for Galaxy	0:92636934a189	Assembly	edward-kirton
agile_wrapper	Quickly match reads to a reference genome or sequence file	0:d6a426afaa46	Next Gen Mappers, Sequence Analysis	simonl
asdf	asdf	-1:0000000000000	Statistics, Text Manipulation	vivek
assemblstats	Summarise an assembly (e.g. N50 metrics)	0:6544228ea290	Next Gen Mappers, Sequence Analysis	konradpaszkiewicz
bam_to_bigwig	Generate BigWig coverage files from BAM files. Allows gapped reads to be split (useful for RNA-Seq). Calculates	5:5b40b93ebae3	Convert Formats, SAM, Visualization	lparsons

As Open Source Software: Local Galaxy Instances

- Galaxy is designed for local installation and customization
- Easily integrate new tools
- Easy to deploy and manage on nearly any (unix) system
- Run jobs on existing compute clusters
- Requires a **computational resource** on which to be deployed

<http://getgalaxy.org>

Galaxy is available ...

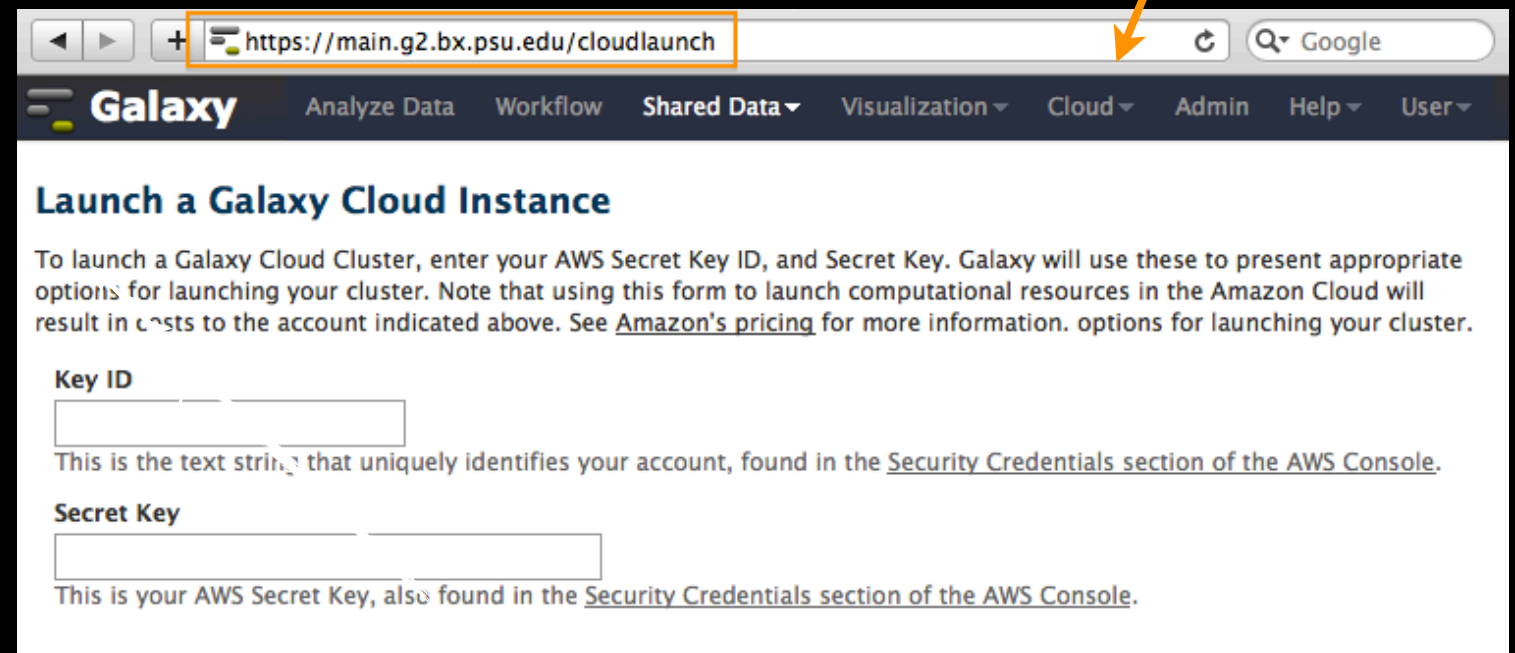
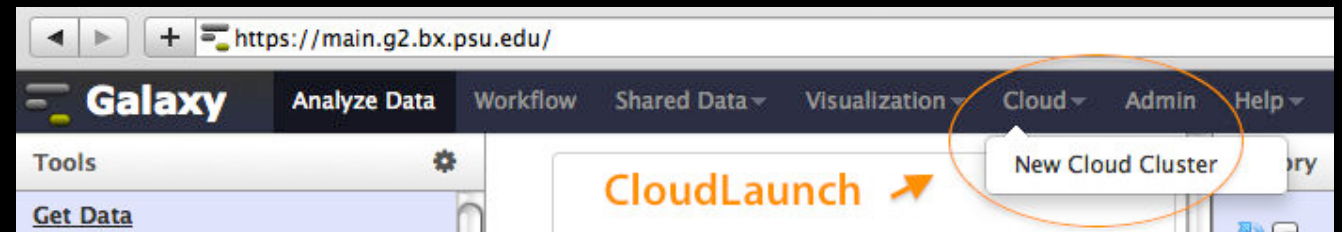
1. Start with a **fully configured and populated** (tools and data) Galaxy instance.
2. Allows you to scale up and down your compute assets as needed.
3. Someone else manages the data center.

- **On the Cloud**

<http://usegalaxy.org/cloud>

Galaxy “CloudMan”

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



Galaxy “Public” Instances

<http://bit.ly/gxyServers>

Interested in:

ChIP-chip and ChIP-seq?

✓ Cistrome

Statistical Analysis?

✓ Genomic Hyperbrowser

Protein synthesis?

✓ GWIPS-viz

de novo assembly?

✓ CBIIT Galaxy

Reasoning with ontologies?

✓ OPPL Galaxy

Repeats!

✓ RepeatExplorer

Everything?

✓ Andromeda

Plus many more

Galaxy Resources and Community

Mailing Lists (very active)

Unified Search

Issues Board

Events Calendar, News Feed

Community Wiki

GalaxyAdmins

Screencasts

Tool Shed

Public Installs

CiteULike group, Mendeley mirror

Annual Community Meeting

<http://wiki.galaxyproject.org>

Galaxy Resources and Community: Mailing Lists

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

Galaxy-Announce

Project announcements, low volume, moderated

Low volume (42 posts in 2012, 2100+ members)

Galaxy-User

Questions about using Galaxy and usegalaxy.org

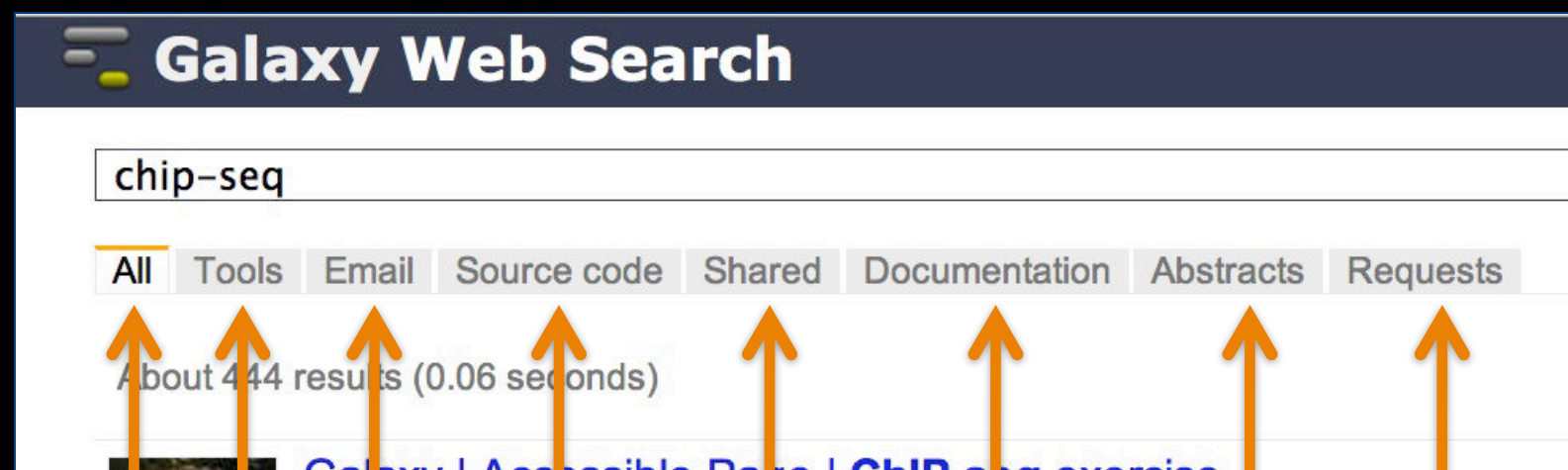
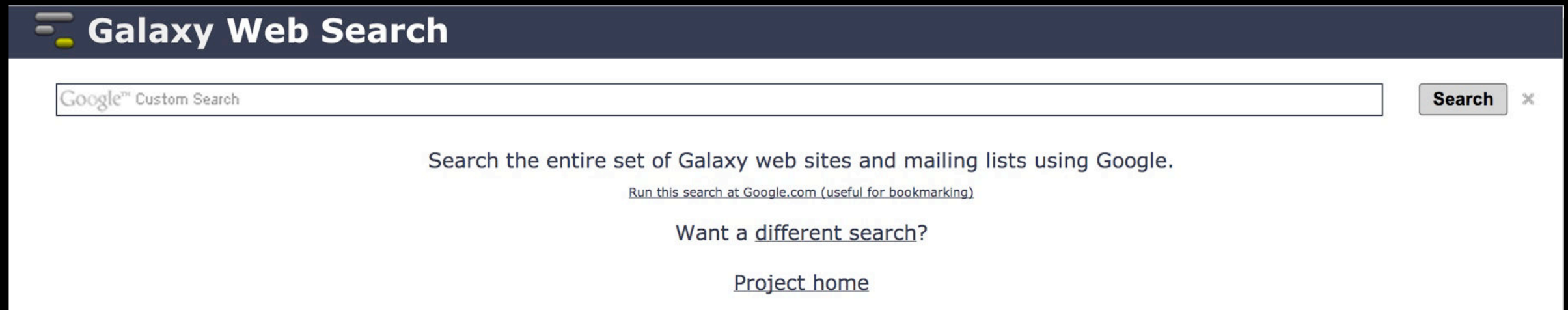
High volume (2900 posts in 2012, 2700+ members)

Galaxy-Dev

Questions about developing for and deploying Galaxy

High volume (4500 posts in 2012, 900+ members)

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

Common to all Development contributors and general users, the Trello Issue Board replaced bitbucket in 2012:

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

The screenshot displays the Trello web interface for the 'Galaxy: Development Inbox' board. The board is structured into four columns: 'Inbox', 'Developer ideas', 'Bug Reports', and 'Issues from Bitbucket'. The 'Inbox' column is currently empty. The 'Developer ideas' column contains three cards: 'Google Drive / Dropbox / Box / ... integration', 'Standalone web application(s) for visualizations', and 'Assistive UI'. The 'Bug Reports' column contains two cards: '823: picard index indicates failure, but it is successful' and '822: cannot run updatencbi.sh'. The 'Issues from Bitbucket' column contains four cards: '5: Option to disable automatic history creation', '6: Option to require that histories have names', '8: More flexible output handlers', and '10: Allow overriding parameters when running a workflow'. The right sidebar features a 'Members' section with a grid of 12 user avatars, a 'Board' section with 'Options', 'Add List', and 'Search and Filter Cards' buttons, and an 'Activity' section. The 'Activity' section is highlighted with an orange border and shows a list of recent actions, including 'Dannon Baker enabled self join on this board. yesterday at 8:35am' and 'Dannon Baker moved Change in # of'. The browser address bar at the top shows the URL 'https://trello.com/board/galaxy-development-inbox/50686d0302dfa79d13d90c45'.

Galaxy Wiki

FrontPage

Locked History Actions

Galaxy

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

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

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!](#)

Galaxy Community Conference 2013

30 June - 2 July

OSLO


University of Oslo

Poster abstracts due 3 May

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](#)

Wiki

[Help](#) • [All Pages](#)

Monday, July 1, 13

32

Events

News

Galaxy Wiki

Login | Search:

Titles



Events

Locked

History






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](#) .

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.

Upcoming Events



Date	Topic/Event	Venue/Location	Contact
April 29 - May 2	Introduction to Galaxy Workshops 2013 Galaxy Workshop Tour	Washington University in St. Louis	Dave Clements
		Saint Louis University	
		University of Missouri Columbia	
May 14-16	Tutorial: <i>Exploring and Enabling Biomedical Data Analysis with Galaxy</i>	Great Lakes Bioinformatics Conference (GLBIO) 2013, Pittsburgh, Pennsylvania, United States	Anton Nekrutenko
May 15	GalaxyAdmins May 2013 Meetup	GalaxyAdmins web meetup	Srinivas Maddhi, Dave Clements
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	James Taylor
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	Sandra Dèrozier, Valentin Loux, Véronique Martin <veronique.martin AT jouy DOT inra DOT fr>
May 22 May 30	Analyse de données issues de séquenceurs nouvelle génération sous Galaxy Les deux ateliers sont maintenant complets		Jean-François Gibrat, Valentin Loux, Véronique Martin <veronique.martin AT jouy DOT inra DOT fr>
May 24 June 19	Introduction to Galaxy A Genomics Virtual Lab for Cancer Research		Nikhil Joshi <najoshi AT ucdavis DOT edu> Dominique Gorse

Galaxy Wiki

Login | Search:

Titles

Text

News

Locked

History

Actions

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](#) .

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

- Galaxy News Briefs
- Galaxy Updates
- Galaxy on Twitter
- Events
- Learn
- Support
- About the Galaxy Project

News Items

Environmental Metabolomics + Galaxy

A new UK-China collaboration in environmental metabolomics between the University of Birmingham, BGI and *GigaScience* has received funding from the UK's Natural Environment Research Council (NERC).

The first metabolomics project will send a developer from the University of Birmingham's School of Biosciences, to Hong Kong to work with *GigaScience* personnel on extending Galaxy for use in metabolomics data analyses.

"Metabolomics involves the detection and quantification of small molecules (metabolites) in living organisms and can provide an indication of their cellular condition and health. The toxicological responses of organisms to pollutants can be studied using environmental metabolomics, enabling researchers to discover diagnostic markers for monitoring and risk assessment of our environment. Research at Birmingham focuses extensively on the metabolic responses of the freshwater model organism, *Daphnia*, to both pollutants and engineered nanomaterials."

See the [official announcement](#) for more details.

Peter Li
GigaScience

Posted to the [Galaxy News](#) on 2013-04-22

Galaxy @ ASMS 2013

Galaxy will have a significant presence at the 61st ASMS Conference on Mass Spectrometry and Allied Topics being held in Minneapolis, Minnesota, June 9-13. Galaxy related content includes the *Galaxy Framework as a Solution for MS-based Informatics* workshop and at least 9 posters either directly about or using Galaxy.

If you do research in proteomics than please consider attending.

Dave Clements

Posted to the [Galaxy News](#) on 2013-04-19

April 8, 2013 Galaxy Security Release

Galaxy Community Conference 2013

30 June - 2 July
University of Oslo

Poster abstracts due 3 May

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

Wiki

Help • All Pages
Recent Changes 
Search • Create Page



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



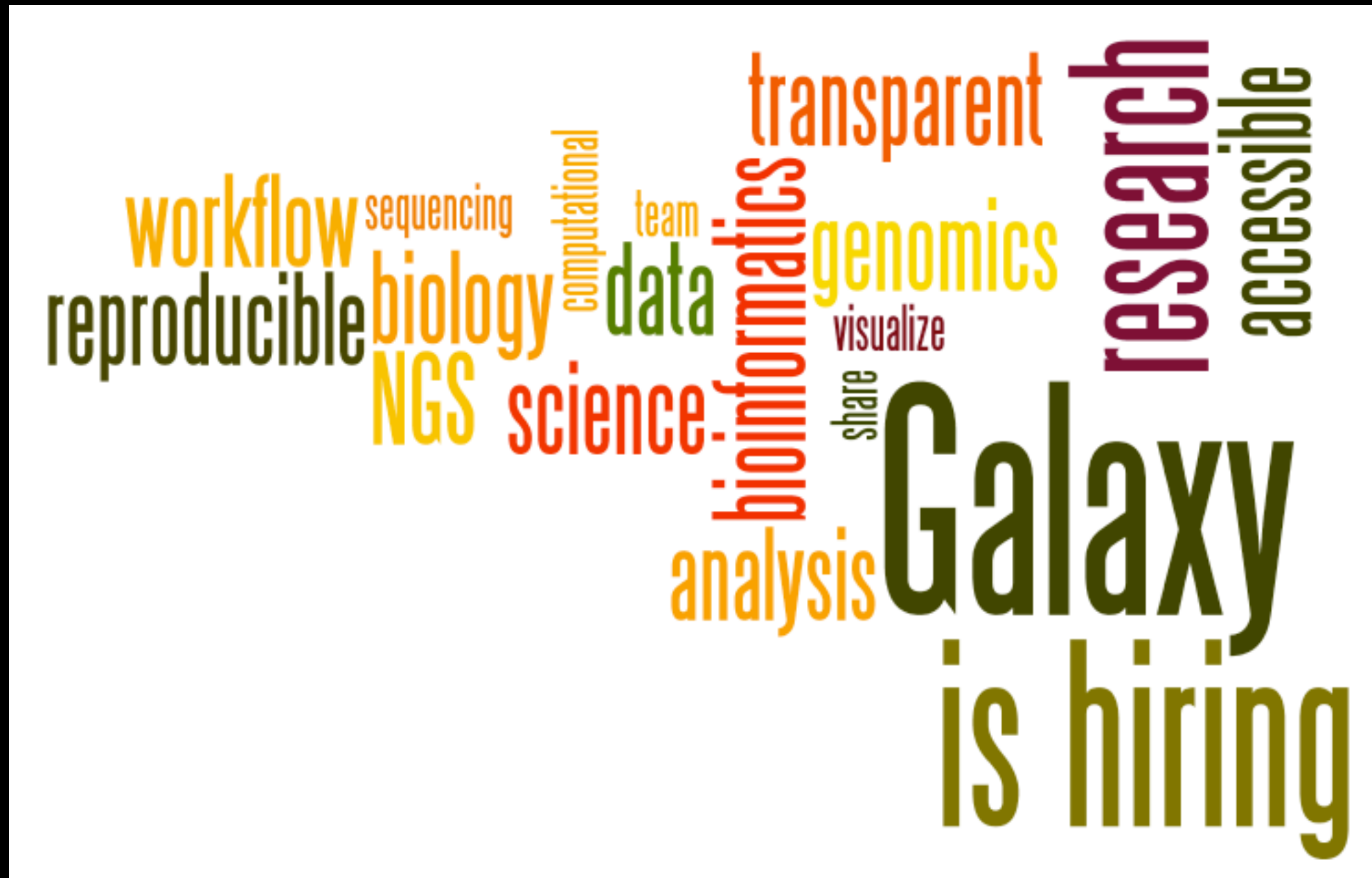
Anton Nekrutenko



James Taylor

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

Galaxy is hiring post-docs and software engineers
at both Emory and Penn State.



Please help.

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

Acknowledgements

Dave Clements
Dannon Baker
Enis Afgan

GCC 2013 Organizing Committee
The Galaxy Team
You!

University of Oslo

AWS Education Grant

NIH NSF Huck Institute
Penn State University Emory University