Galaxy @ UCSC

March 13, 2013

Jennifer Hillman-Jackson
Penn State University

galaxyproject.org
genome.ucsc.edu

Agenda

Galaxy project mission

Who’s on the team

Overview & Terminology

Graphical Example - 101

Enough now ... let’s see it!
- Wiki, Tools, Histories, Sharing, Workflows, etc.

Agenda

Galaxy project mission

Who’s on the team

Overview & Terminology

Graphical Example - 101

Enough now ... let’s see it!
- Wiki, Tools, Histories, Sharing, Workflows, etc.

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)

Agenda

Galaxy project mission

Who’s on the team

Overview & Terminology

Graphical Example - 101

Enough now ... let’s see it!
- Wiki, Tools, Histories, Sharing, Workflows, etc.

Enis Afgan

Dannon Baker

Dan Blankenberg

Dave Bouvier

Dave Clements

Nate Coraor

Carl Eberhard

Jeremy Goecks

Sam Guerler

Jen Jackson

Greg von Kuster

Ross Lazarus

Anton Nekrutenko

James Taylor

?

The Galaxy Team

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

Agenda

Galaxy project mission

Who’s on the team

Overview & Terminology

Graphical Example - 101

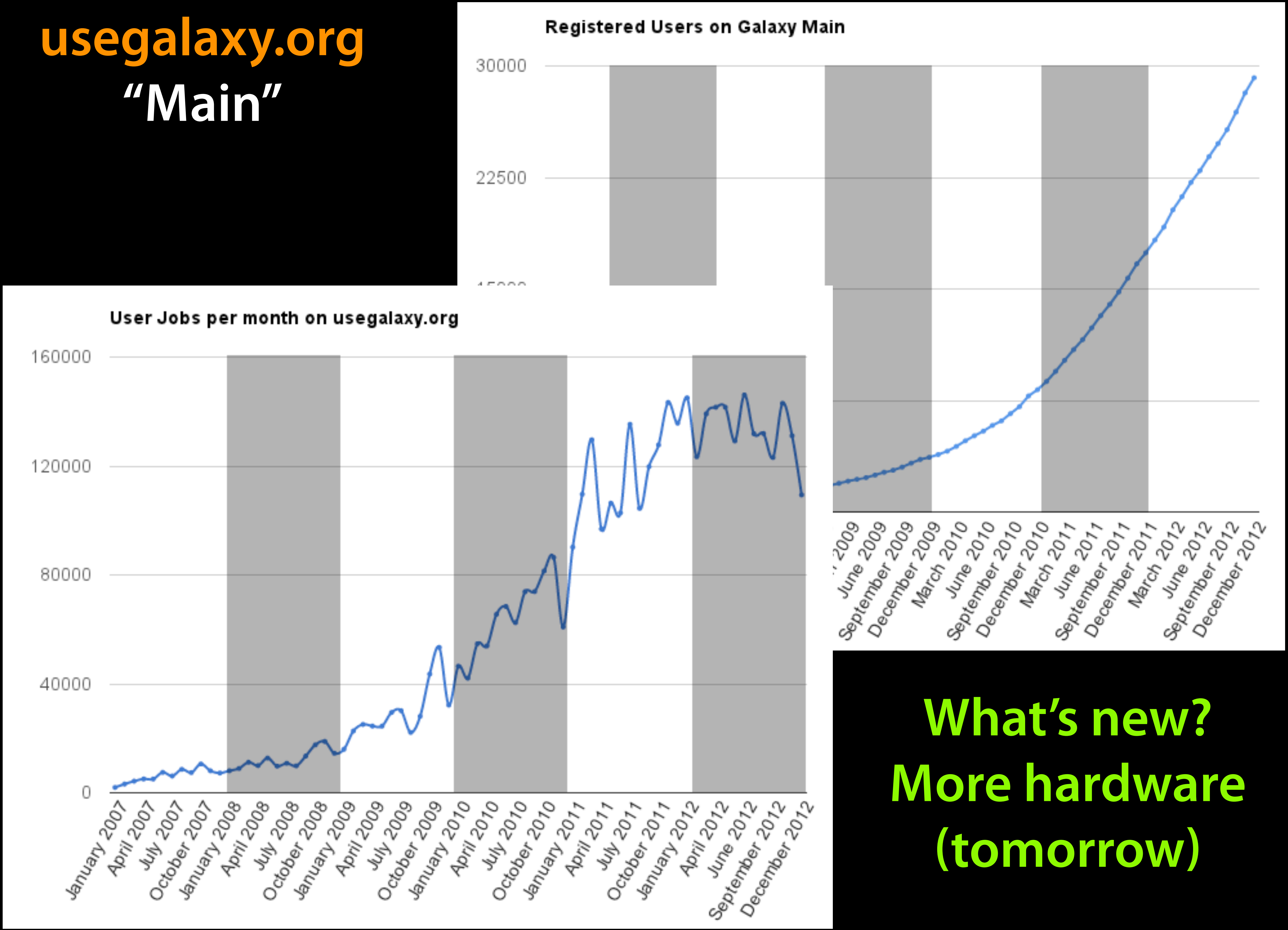
Enough now ... let’s see it!
- Wiki, Tools, Histories, Sharing, Workflows, etc.

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

usegalaxy.org
“Main”



What’s new?
More hardware
(tomorrow)

getgalaxy.org
“Local”

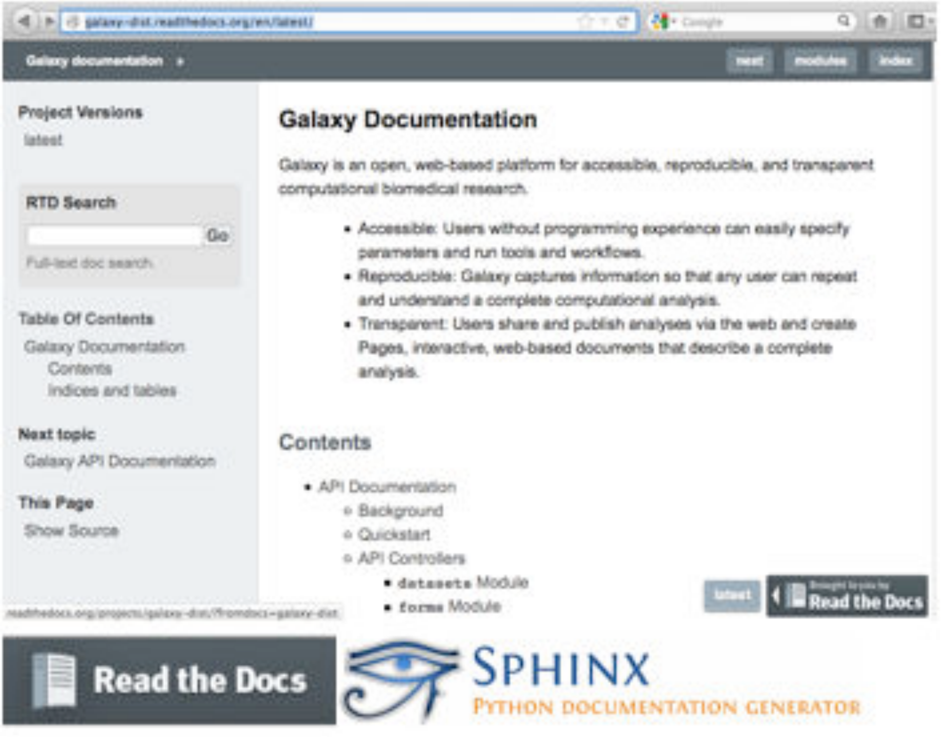
wiki.galaxyproject.org/DevNewsBriefs

galaxy-dist.readthedocs.org

bitbucket.org/galaxy/galaxy-dist

Code Downloads -- how often??

Alas, this information does not appear to be available from Bitbucket. Therefore, *we don't know.*


galaxy-dist.readthedocs.org/en/latest/
Google
Galaxy documentation »
next
modules
index
Go
latest
Brought to you by Read the Docs
readthedocs.org/projects/galaxy-dist/?fromdocs=galaxy-dist
Read the Docs
SPHINX
PYTHON DOCUMENTATION GENERATOR

Project Versions
latest

RTD Search
Full-text doc search.

Table Of Contents
Galaxy Documentation
Contents
Indices and tables

Next topic
Galaxy API Documentation

This Page
Show Source

Galaxy Documentation

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.

Contents

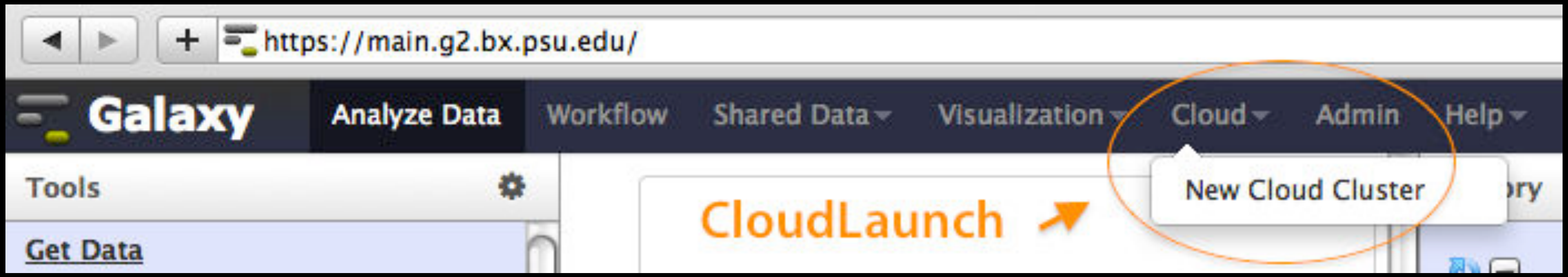
- API Documentation
  - Background
  - Quickstart
  - API Controllers
    - `datasets` Module
    - `forms` Module

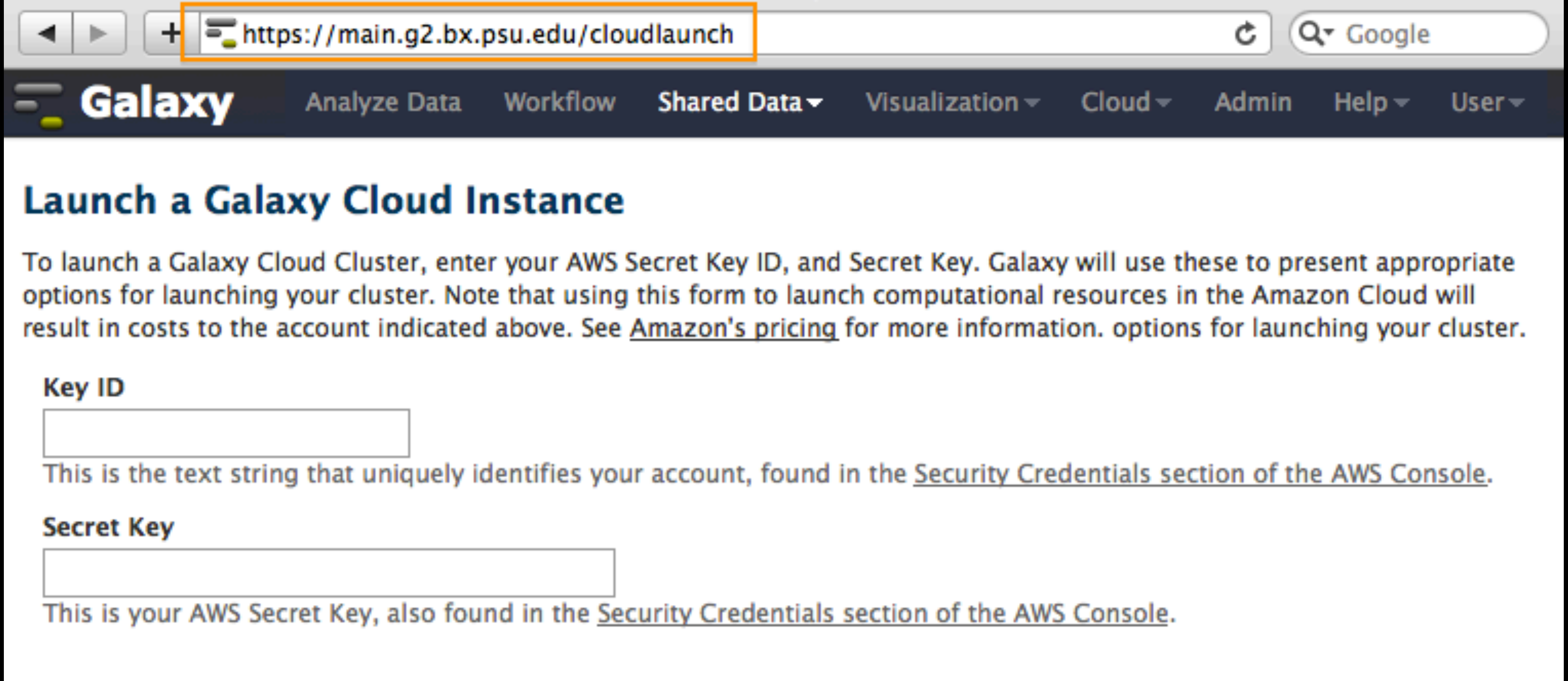
What’s new?

- “readthedocs” documentation

usegalaxy.org/cloud

“CloudMan”





Launch a Galaxy Cloud Instance

To launch a Galaxy Cloud Cluster, enter your AWS Secret Key ID, and Secret Key. Galaxy will use these to present appropriate options for launching your cluster. Note that using this form to launch computational resources in the Amazon Cloud will result in costs to the account indicated above. See Amazon's pricing for more information. options for launching your cluster.

Key ID

This is the text string that uniquely identifies your account, found in the Security Credentials section of the AWS Console.

Secret Key

This is your AWS Secret Key, also found in the Security Credentials section of the AWS Console.

What’s new?

- Educational grants for cloud time from Amazon
- CloudLaunch directly from within “Main”
- Publications integrating Galaxy cloud workflows

Galaxy CloudMan

http://usegalaxy.org/cloud

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

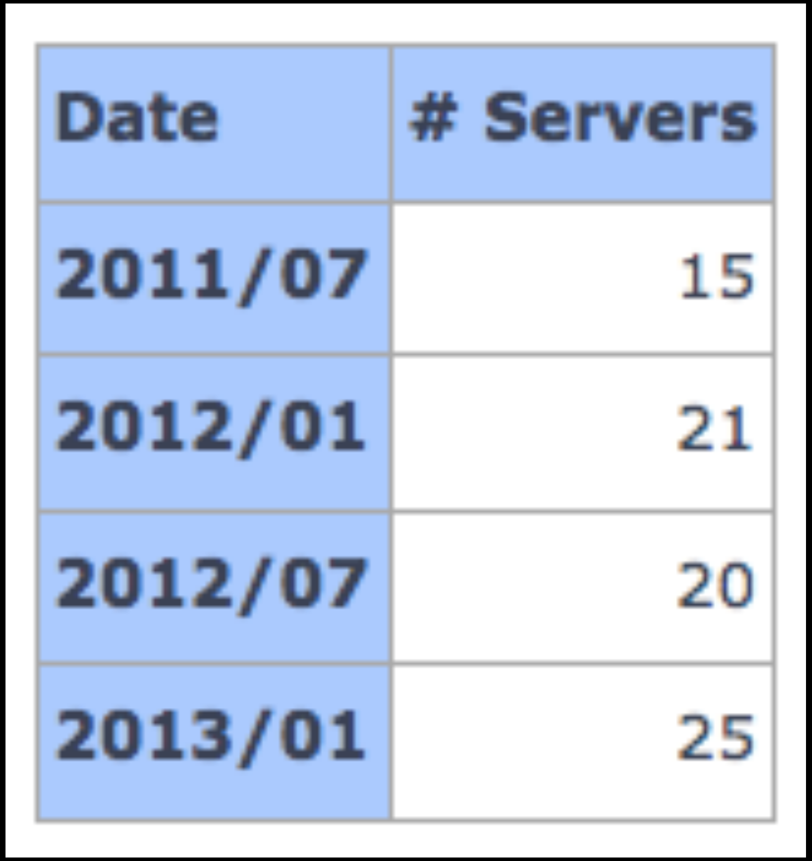
amazon
webservices™

http://aws.amazon.com/education

wiki.galaxyproject.org/PublicGalaxyServers

“Known Publicly Accessible Servers”

This is not an absolute count, but it is a rough measure of the trend.



Date	# Servers
2011/07	15
2012/01	21
2012/07	20
2013/01	25

What’s new?

- “GalaxyAdmins” community group founded

Public Galaxy Instances

http://wiki.galaxyproject.org/PublicGalaxyServers

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

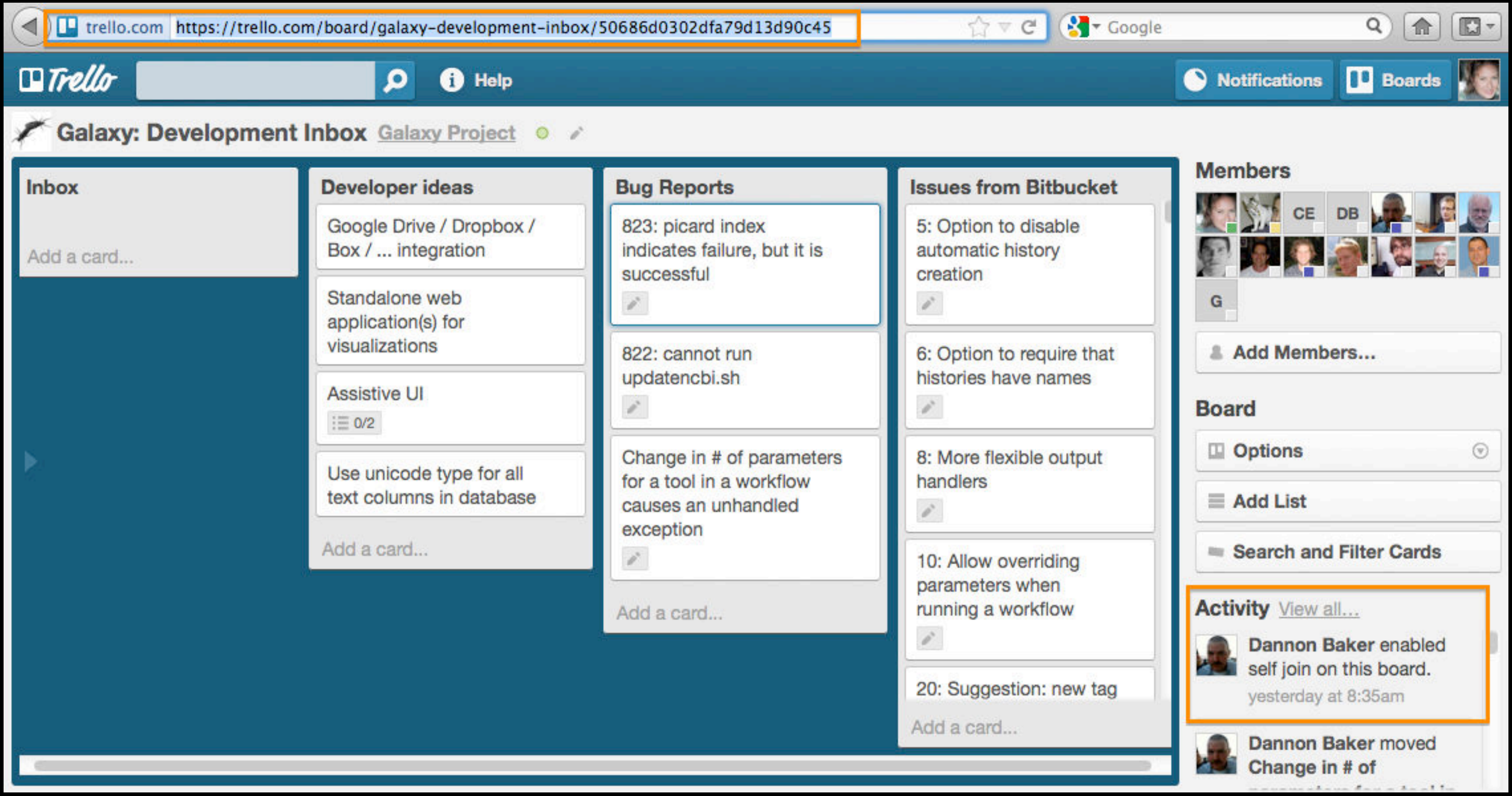
Repeats!
- ✓ RepeatExplorer

Everything?
- ✓ Andromeda

Plus many more

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

http://wiki.galaxyproject.org/Issues



trello.com https://trello.com/board/galaxy-development-inbox/50686d0302dfa79d13d90c45

Galaxy: Development Inbox Galaxy Project

Inbox

Add a card...

Developer ideas

- Google Drive / Dropbox / Box / ... integration
- Standalone web application(s) for visualizations
- Assistive UI 0/2
- Use unicode type for all text columns in database

Add a card...

Bug Reports

- 823: picard index indicates failure, but it is successful
- 822: cannot run updatencbi.sh
- Change in # of parameters for a tool in a workflow causes an unhandled exception

Add a card...

Issues from Bitbucket

- 5: Option to disable automatic history creation
- 6: Option to require that histories have names
- 8: More flexible output handlers
- 10: Allow overriding parameters when running a workflow
- 20: Suggestion: new tag

Add a card...

Members

Board

Activity View all...

Dannon Baker enabled self join on this board.
yesterday at 8:35am

Dannon Baker moved Change in # of

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)

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 tools and datatypes

Data and Tools - new in 2012/2013

Reference Genomes:

Dozens of full genomes added and over a hundred genomes had some content (index, liftOver) added in 2012. New data early 2013 is including **Bowtie2** indexes both on **Main** and **rsync** download area.

- The rsync area was new in 2012, too:

http://wiki.galaxyproject.org/Admin/Data%20Integration

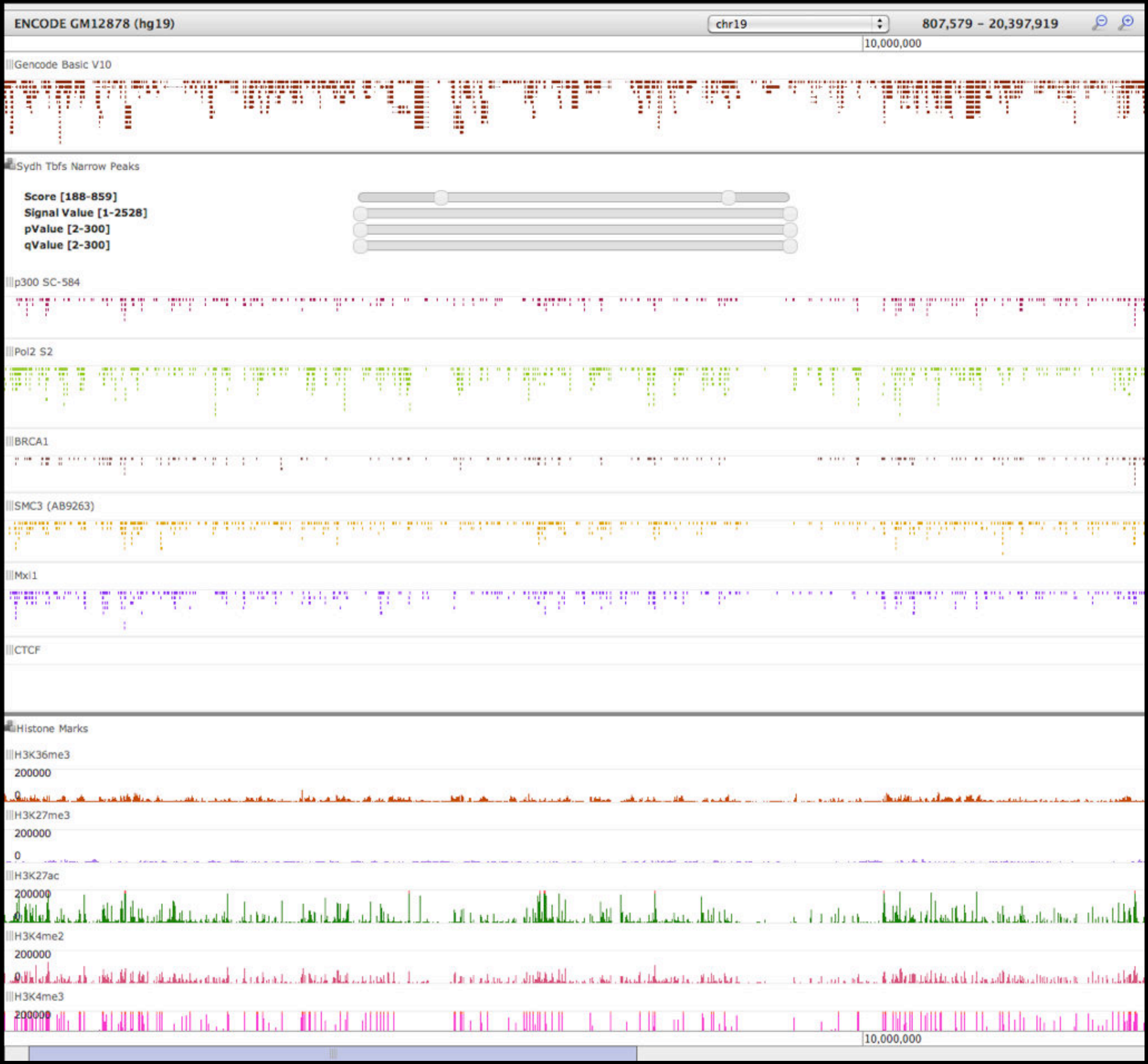
Key Tools Included:

GATK (beta); Updates to the RNA-seq tool set **Bowtie2/Tophat2**, **Cufflinks/merge/diff**; **FreeBayes**; **Trinity** (Tool Shed); **Wormbase** 2; **IGB**; **GenomeSpace**; **Megablast** to use BLAST+; **MPileup**, and the **Tool Factory** (Tool Shed)**:**

"Creating re-usable tools from scripts: The Galaxy Tool Factory," Ross Lazarus, Antony Kaspi, Mark Ziemann, The Galaxy Team, Bioinformatics (28 September 2012)

Visualizations

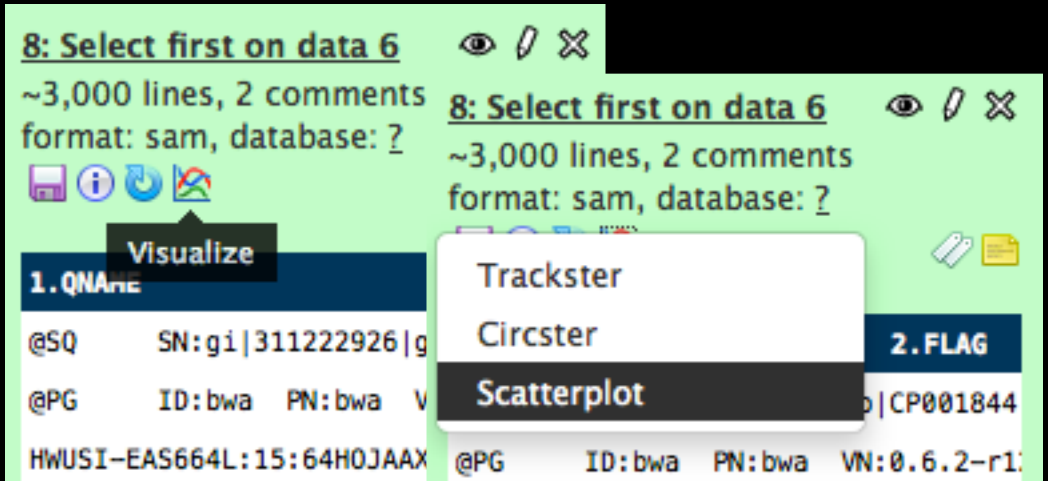
Trackster had significant number of new refinements in 2012 leading to a publication.



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

Visualizations

And Circster (not shown) and Scatterplot were introduced:

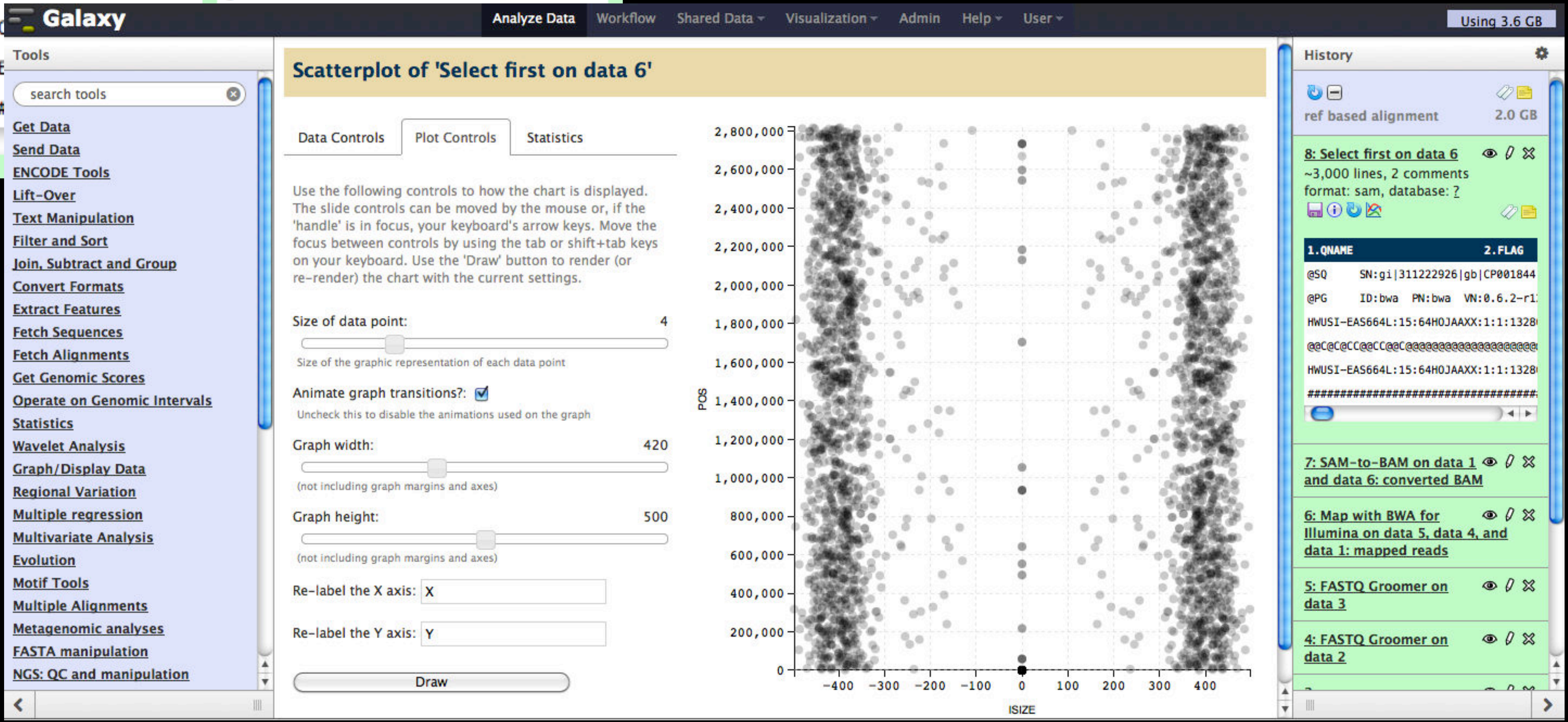


8: Select first on data 6

~3,000 lines, 2 comments

format: sam, database: ?

```
@SQ     SN:gi|311222926|gb|CP001844
@PG     ID:bwa  PN:bwa  VN:0.6.2-r1
HWUSI-EAS664L:15:64HOJAAX
@@C@C@
HWUSI-E
#######
```



Scatterplot of 'Select first on data 6'

Use the following controls to how the chart is displayed. The slide controls can be moved by the mouse or, if the 'handle' is in focus, your keyboard's arrow keys. Move the focus between controls by using the tab or shift+tab keys on your keyboard. Use the 'Draw' button to render (or re-render) the chart with the current settings.

Size of data point: 4

Size of the graphic representation of each data point

Animate graph transitions?: ☑

Uncheck this to disable the animations used on the graph

Graph width: 420

(not including graph margins and axes)

Graph height: 500

(not including graph margins and axes)

Re-label the X axis: X

Re-label the Y axis: Y

8: Select first on data 6

~3,000 lines, 2 comments

format: sam, database: ?

```
@SQ     SN:gi|311222926|gb|CP001844
@PG     ID:bwa  PN:bwa  VN:0.6.2-r1
HWUSI-EAS664L:15:64HOJAAXX:1:1:1328
@@C@C@CC@@CC@@C@@@@@@@@@@@@@@@@@@@@
HWUSI-EAS664L:15:64HOJAAXX:1:1:1328
###################################
```

7: SAM-to-BAM on data 1 and data 6: converted BAM

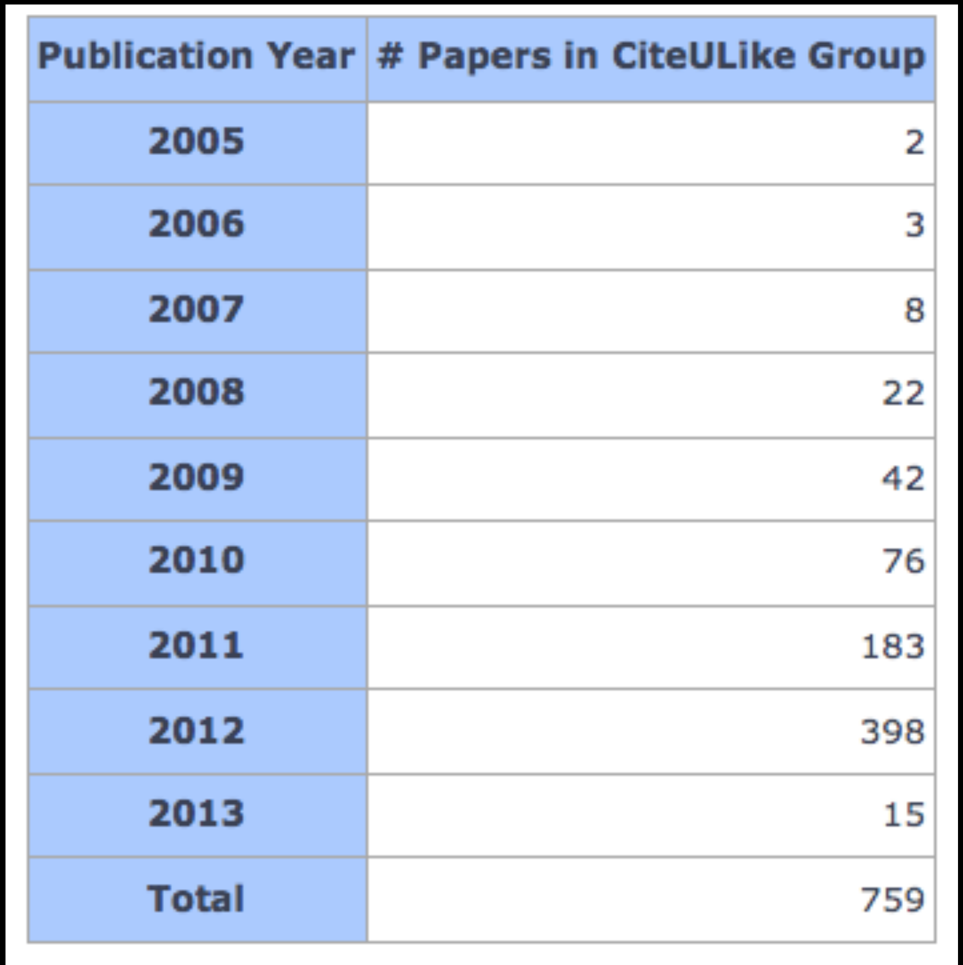
6: Map with BWA for Illumina on data 5, data 4, and data 1: mapped reads

5: FASTQ Groomer on data 3

4: FASTQ Groomer on data 2

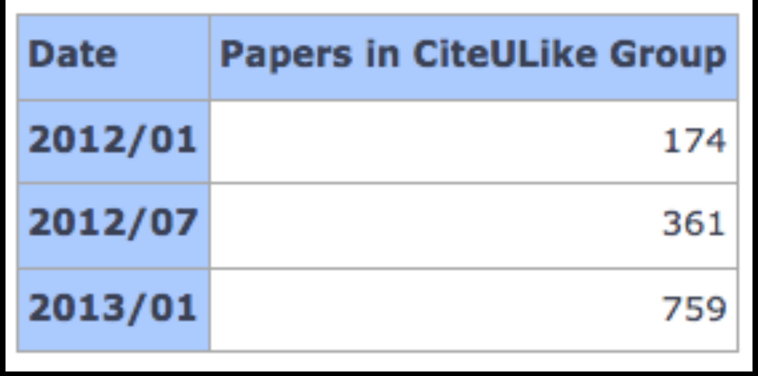
Publications - CiteULike

A Galaxy CiteULike group was started in late 2011. It lists all the pubs that are about, reference, or mention Galaxy that we know about. We started keeping track of this partway through 2011, so it is an undercount for that year (and previous years are almost entirely absent). For years after 2011, it is likely to be more accurate, but still approximate, and to still be an undercount.



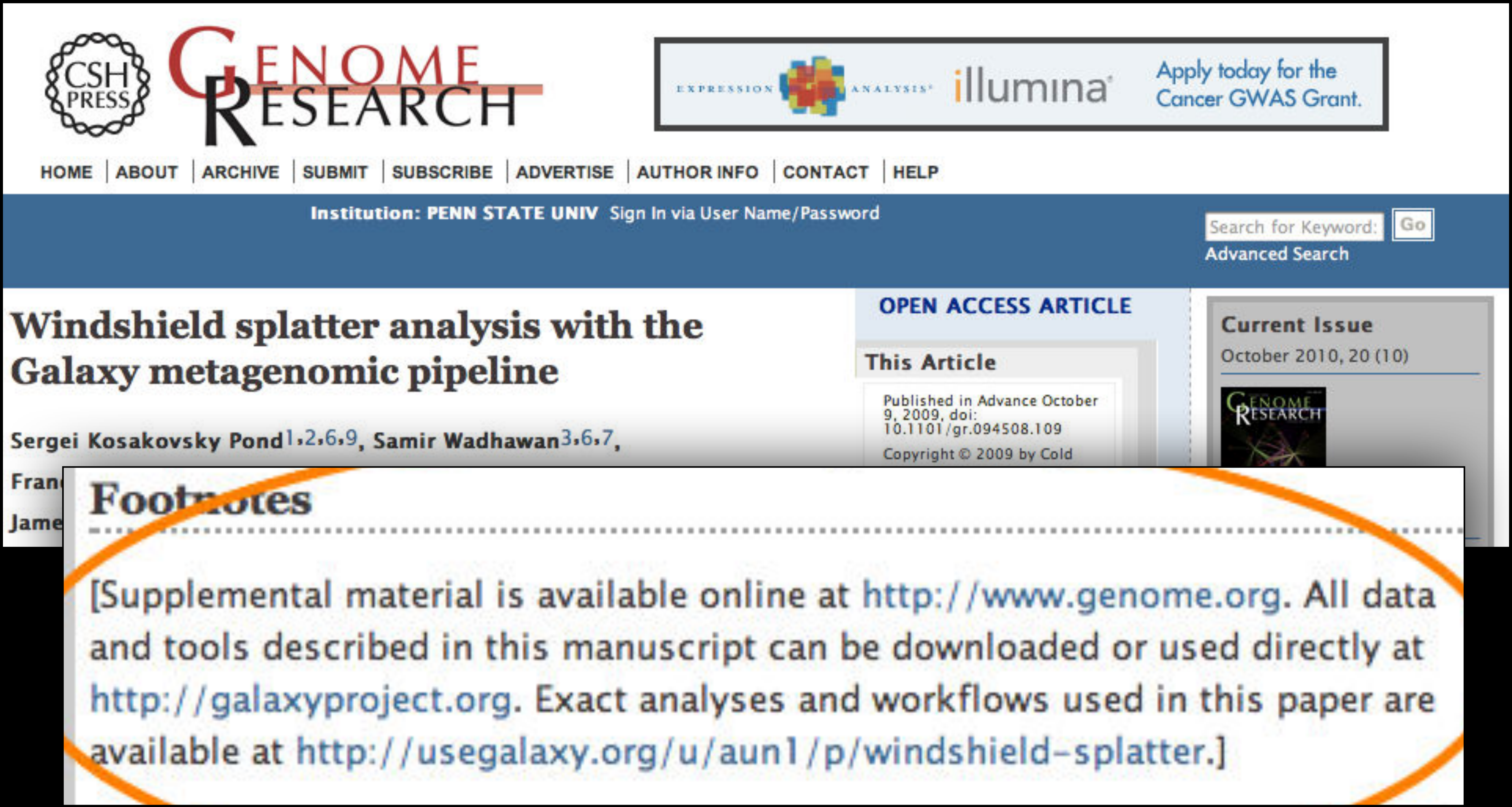
Publication Year	# Papers in CiteULike Group
2005	2
2006	3
2007	8
2008	22
2009	42
2010	76
2011	183
2012	398
2013	15
Total	759

As of January 2013, there are this many papers in the Galaxy CiteULike Group:



Date	Papers in CiteULike Group
2012/01	174
2012/07	361
2013/01	759

Publications - Example 1



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
Current Issue
October 2010, 20 (10)
GENOME RESEARCH

Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond[1,2,6,9], Samir Wadhawan[3,6,7],

Fran

Jame

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

Copyright © 2009 by Cold

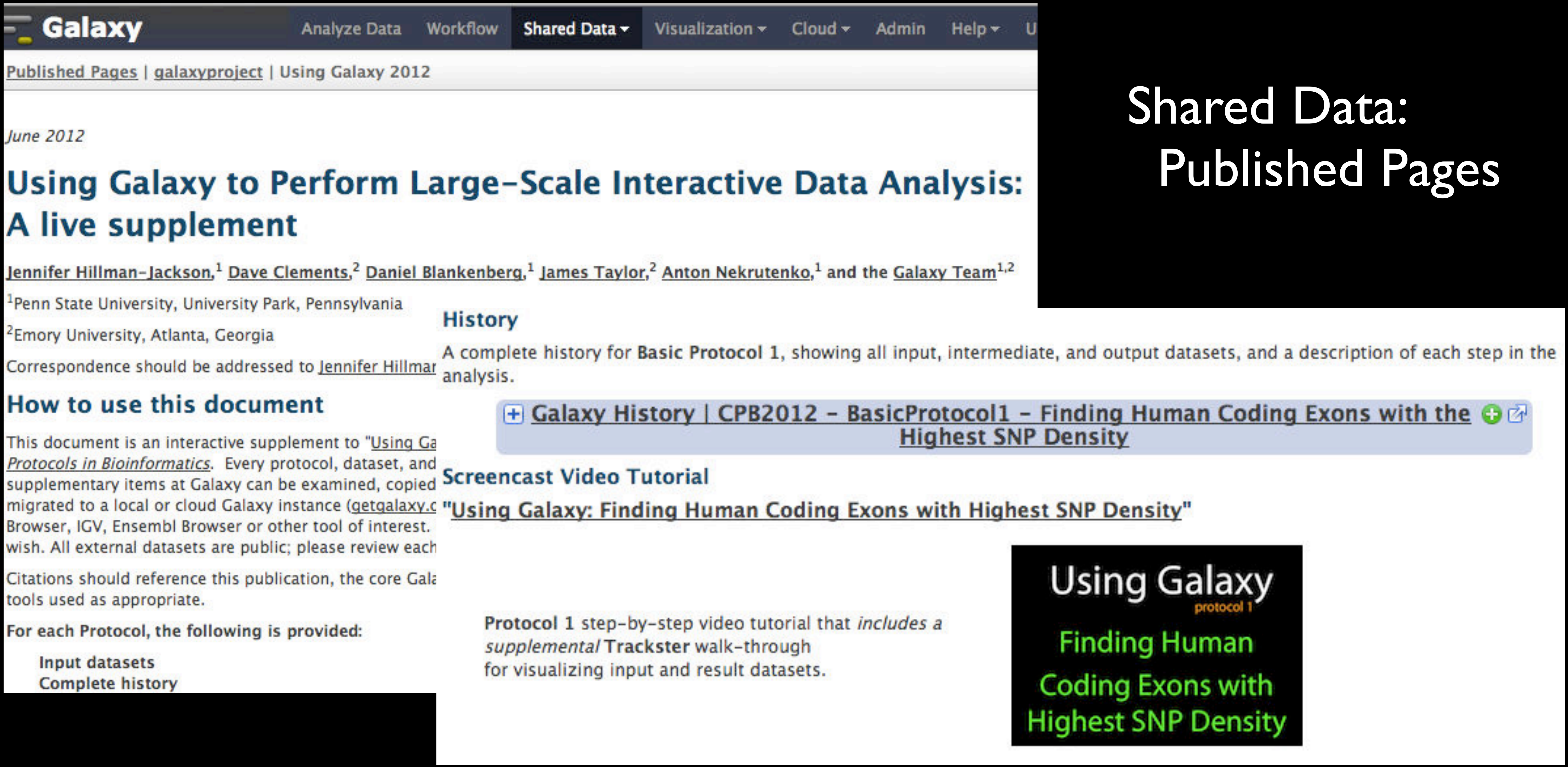
Footnotes

[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.]

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

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

Publications - Example 2



Published Pages | galaxyproject | Using Galaxy 2012

June 2012

Using Galaxy to Perform Large-Scale Interactive Data Analysis: A live supplement

Jennifer Hillman-Jackson,[1] Dave Clements,[2] Daniel Blankenberg,[1] James Taylor,[2] Anton Nekrutenko,[1] and the Galaxy Team[1,2]

[1]Penn State University, University Park, Pennsylvania

[2]Emory University, Atlanta, Georgia

Correspondence should be addressed to Jennifer Hillmar

How to use this document

This document is an interactive supplement to "Using Ga *Protocols in Bioinformatics*. Every protocol, dataset, and supplementary items at Galaxy can be examined, copied migrated to a local or cloud Galaxy instance (getgalaxy.c Browser, IGV, Ensembl Browser or other tool of interest. wish. All external datasets are public; please review each

Citations should reference this publication, the core Gala tools used as appropriate.

For each Protocol, the following is provided:

- **Input datasets**
- **Complete history**

History

A complete history for **Basic Protocol 1**, showing all input, intermediate, and output datasets, and a description of each step in the analysis.

⊞ Galaxy History | CPB2012 - BasicProtocol1 - Finding Human Coding Exons with the Highest SNP Density

Screencast Video Tutorial

"Using Galaxy: Finding Human Coding Exons with Highest SNP Density"

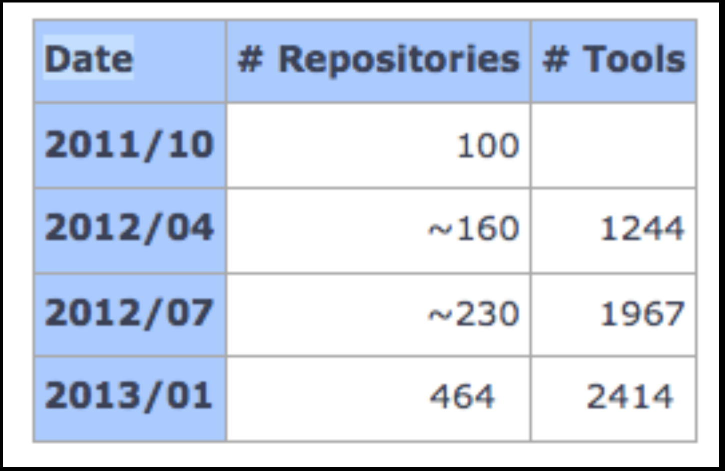
Protocol 1 step-by-step video tutorial that *includes a supplemental* **Trackster** walk-through for visualizing input and result datasets.

Shared Data:
Published Pages

Get all of the data and follow a tutorial start to finish using the supplimental methods, workflows, and screencasts.

http://main.g2.bx.psu.edu/u/galaxyproject/p/using-galaxy-2012

Galaxy Tool Shed



Date	# Repositories	# Tools
2011/10	100	
2012/04	~160	1244
2012/07	~230	1967
2013/01	464	2414

- Allow users to share "containers" of tools, datatypes, workflows, sample data, READMEs, and automated installation scripts for tool dependencies.
- Integration with Galaxy instances to automate tool installation and updates.
- Is currently undergoing an audit to identify "valid tools" and set standards with community collaborators - the "IUC".
- In process of supporting **improved dependency documentation** and upgrading installation processes.

toolshed.g2.bx.psu.edu

Galaxy Tool Shed

- There is a "**Main**" tool shed hosted by the core Galaxy team, but *satellite tools sheds are encouraged*. We'd like to learn about them and list on our wiki, as public Galaxies are.
- This "**Main**" tool shed currently tracks the Galaxy distribution as many enhancements are linked/dependent.
- The complete list of updates, usage examples, features, etc. , are in the News Brief Archives at: *wiki.galaxyproject.org/DevNewsBriefs*
- The tool shed documentation covers a LOT of material relevant to the Galaxy framwork as a whole: *wiki.galaxyproject.org/Tool%20Shed*

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

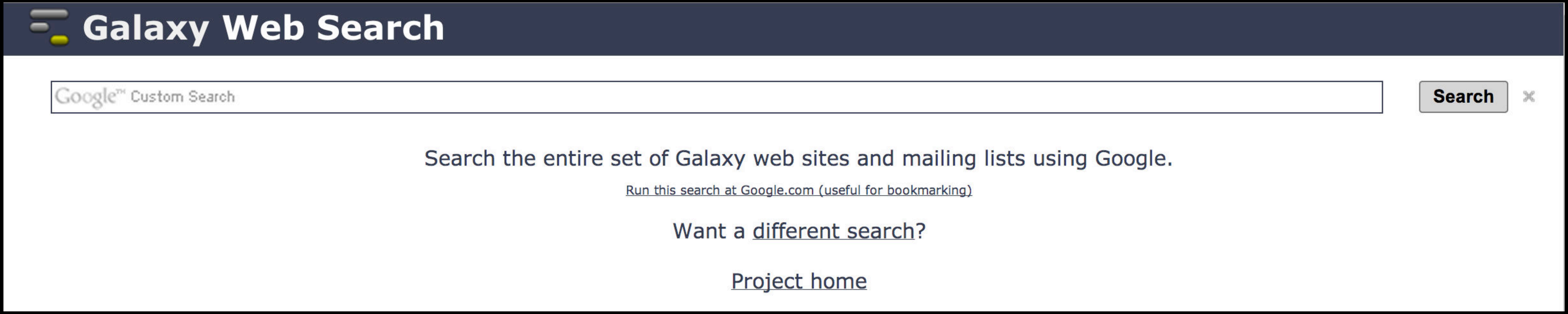
High volume (2900 posts, 2700 members in 2012)

Galaxy-Dev

Questions about developing for and deploying Galaxy

High volume (4500 posts, 850 members in 2012)

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



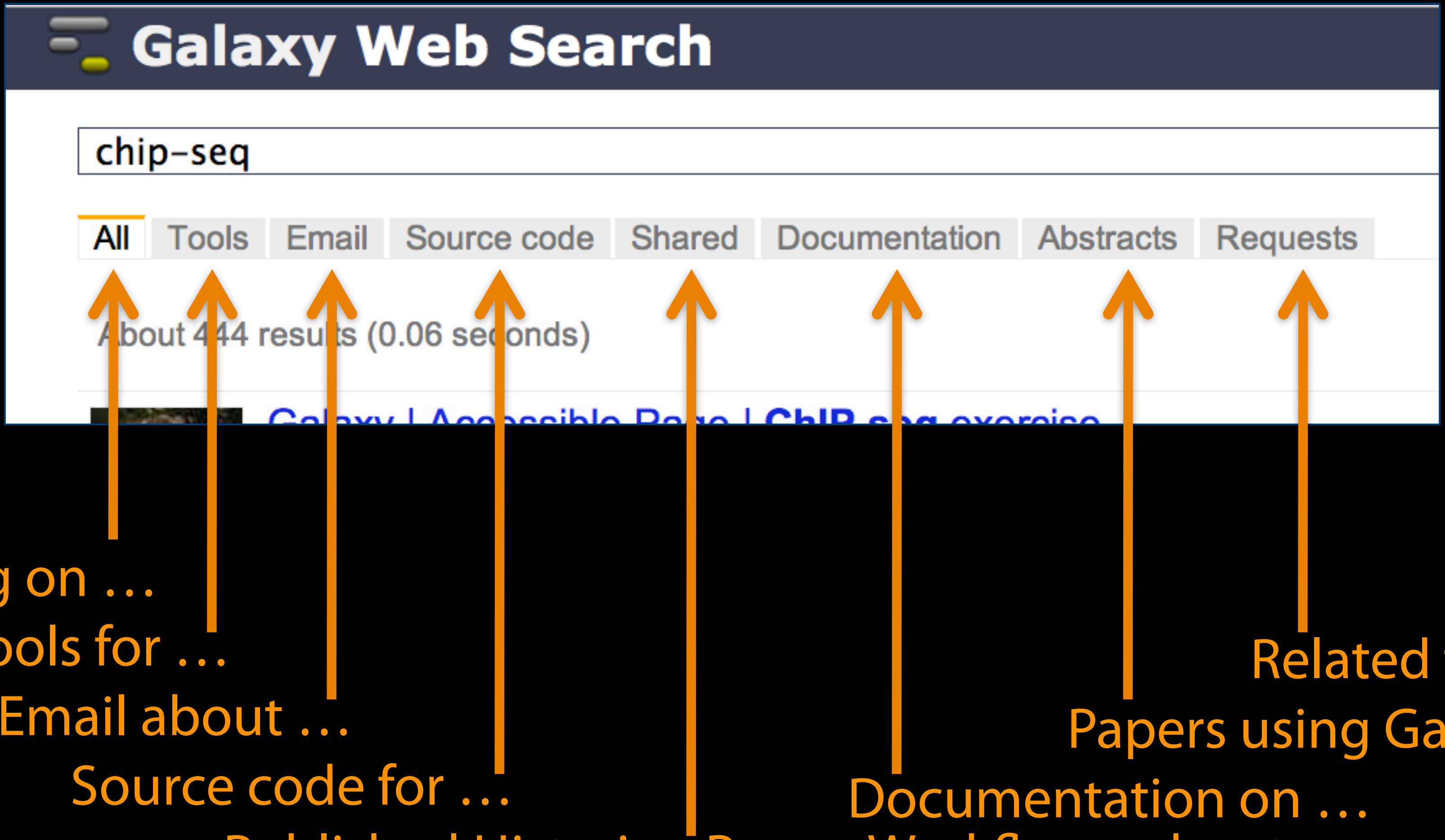
Google™ Custom Search

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



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

Agenda

Galaxy project mission

Who’s on the team

Overview & Terminology

Graphical Example - 101

Enough now ... let’s see it!
- Wiki, Tools, Histories, Sharing, Workflows, etc.

Graphic Example - 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

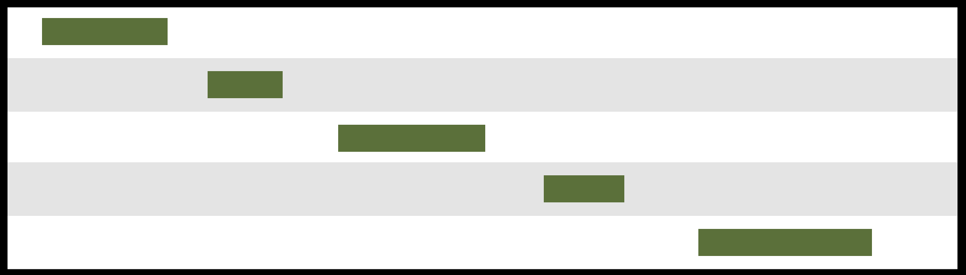
~ http://usegalaxy.org/galaxy101

Exons & Repeats: The General Flow

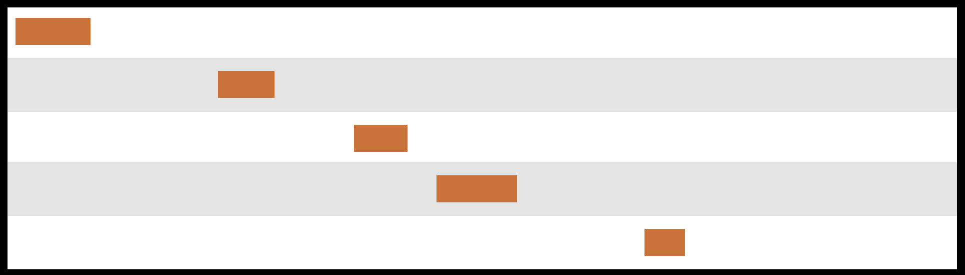
- Get some data
  - Coding exons on chromosome 22
  - Repeats on chromosome 22
- Mess with it
  - Identify which exons have repeats
  - Count repeats per exon
  - Rearrange data into standardized format

For today, we will walk-through initial analysis conceptually, then once in Galaxy, import a similar working history, review the completed tasks, create a workflow, and run it. Along the way exploring tool re-run features, sharing, workflow editing, and more.

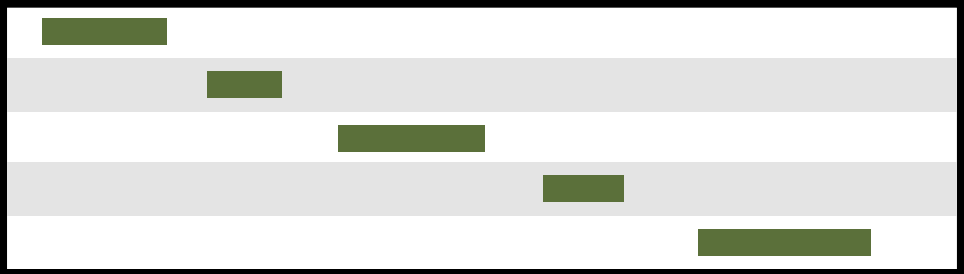
~ http://usegalaxy.org/galaxy101



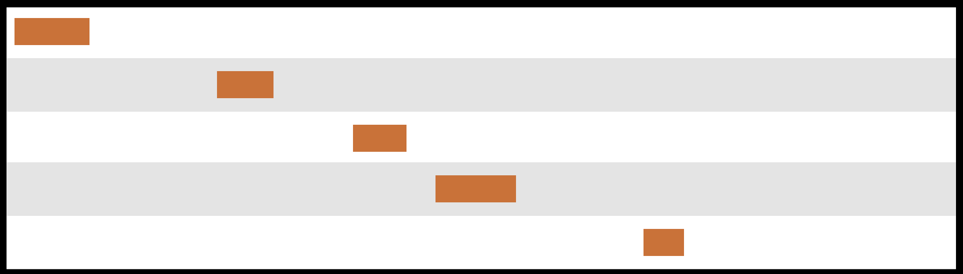
Exons, from UCSC



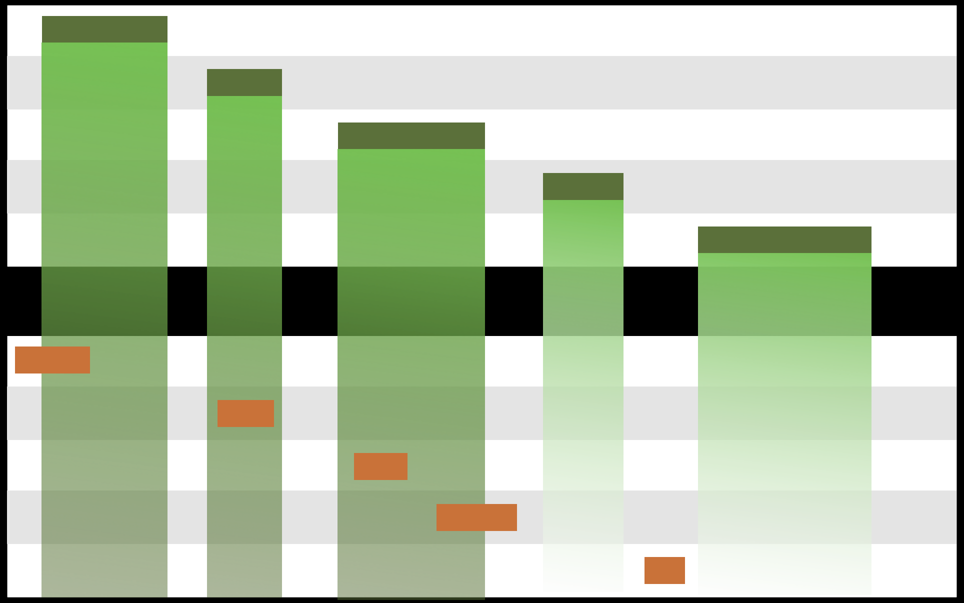
Repeats, from UCSC



Exons, from UCSC



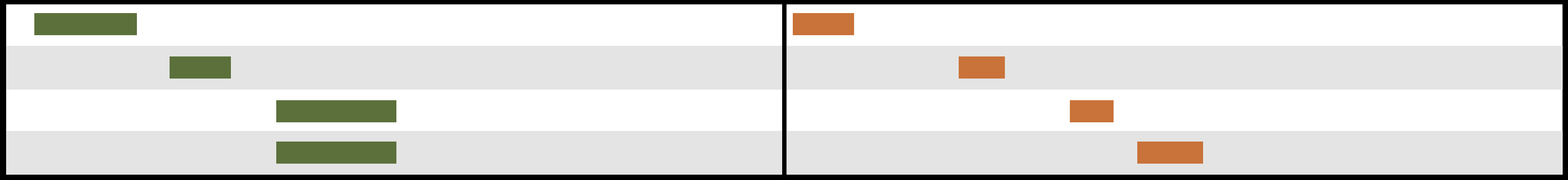
Repeats, from UCSC



Exons, from UCSC

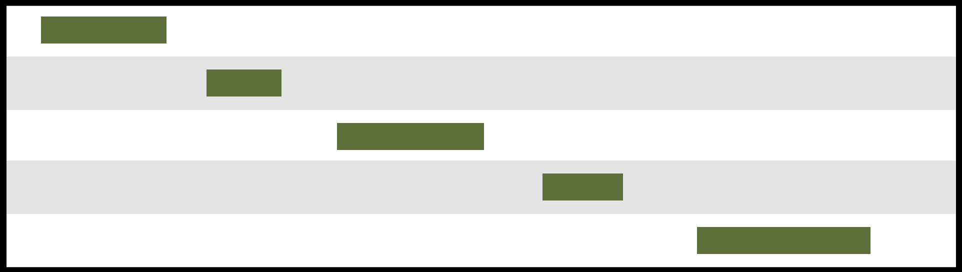
Repeats, from UCSC

Overlap pairings

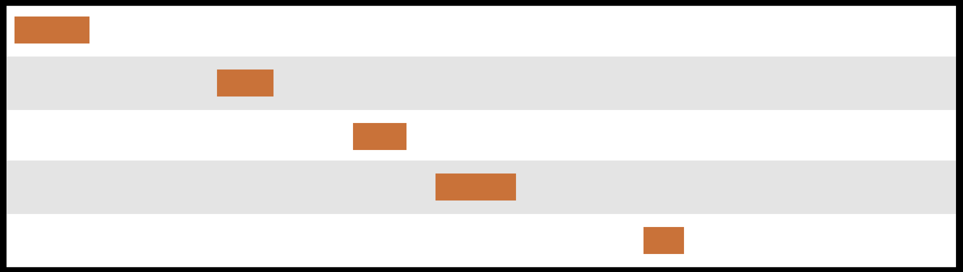


coordinate join

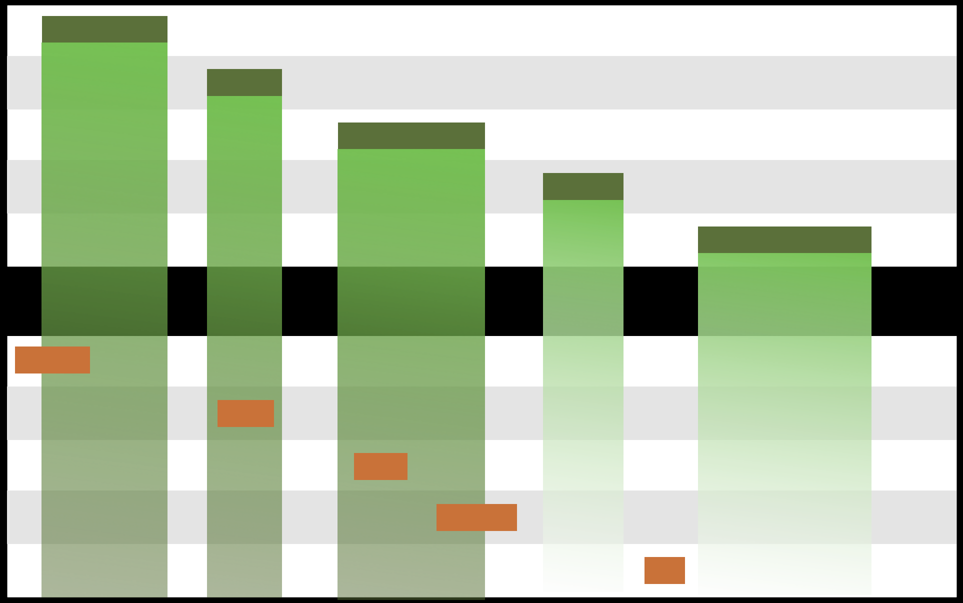
data calculation



Exons, from UCSC



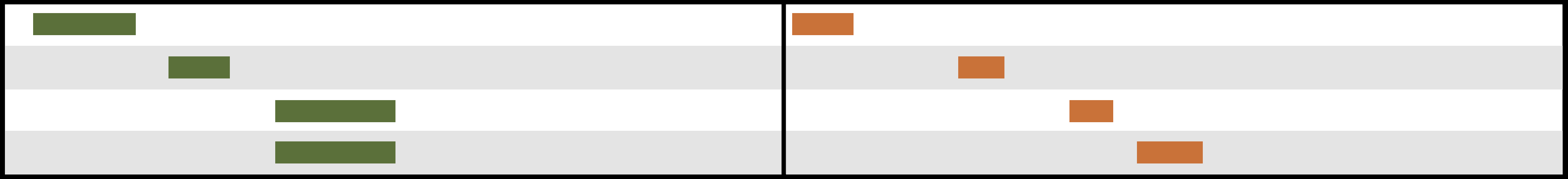
Repeats, from UCSC

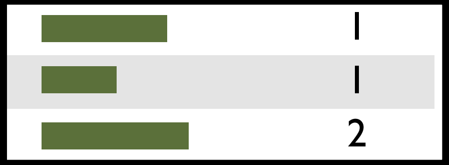


Exons, from UCSC

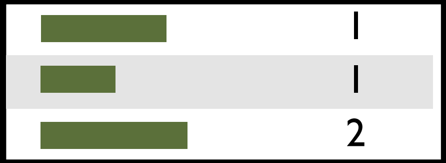
Repeats, from UCSC

Overlap pairings



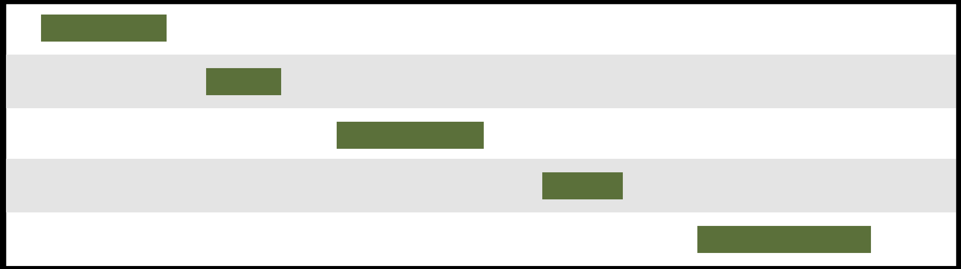


Exon overlap counts

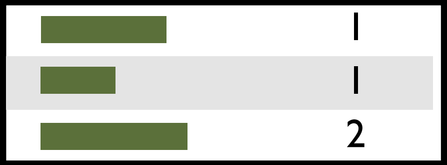


🟩	1
🟩	1
🟩	2

Exon overlap counts

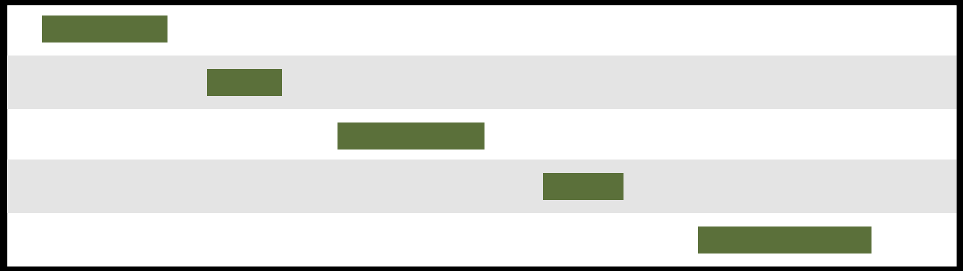


Exons, from UCSC



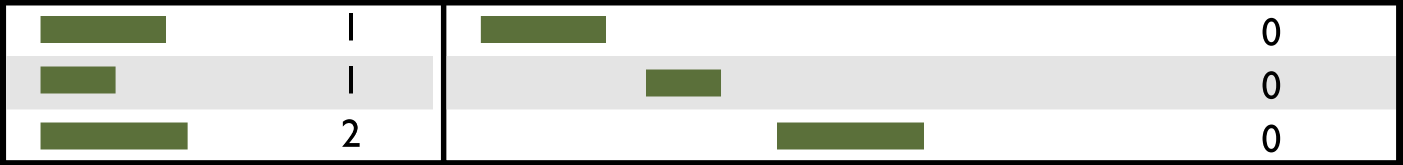
🟩	1
🟩	1
🟩	2

Exon overlap counts



🟩
🟩
🟩
🟩
🟩

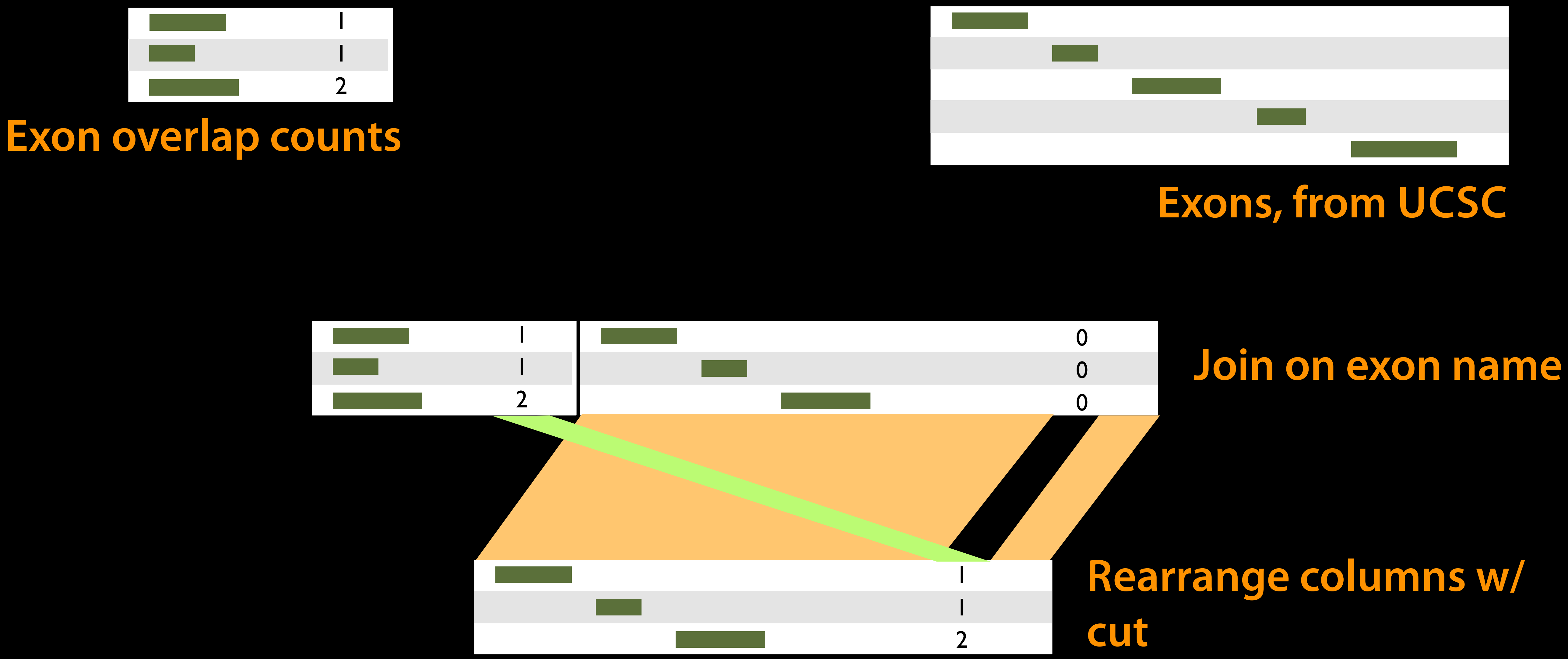
Exons, from UCSC



🟩	1	🟩	0
🟩	1	🟩	0
🟩	2	🟩	0

Join on exon name

relational join



Exon overlap counts

Exons, from UCSC

Join on exon name

Rearrange columns w/ cut

data manipulation

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

- The analysis in the example 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.

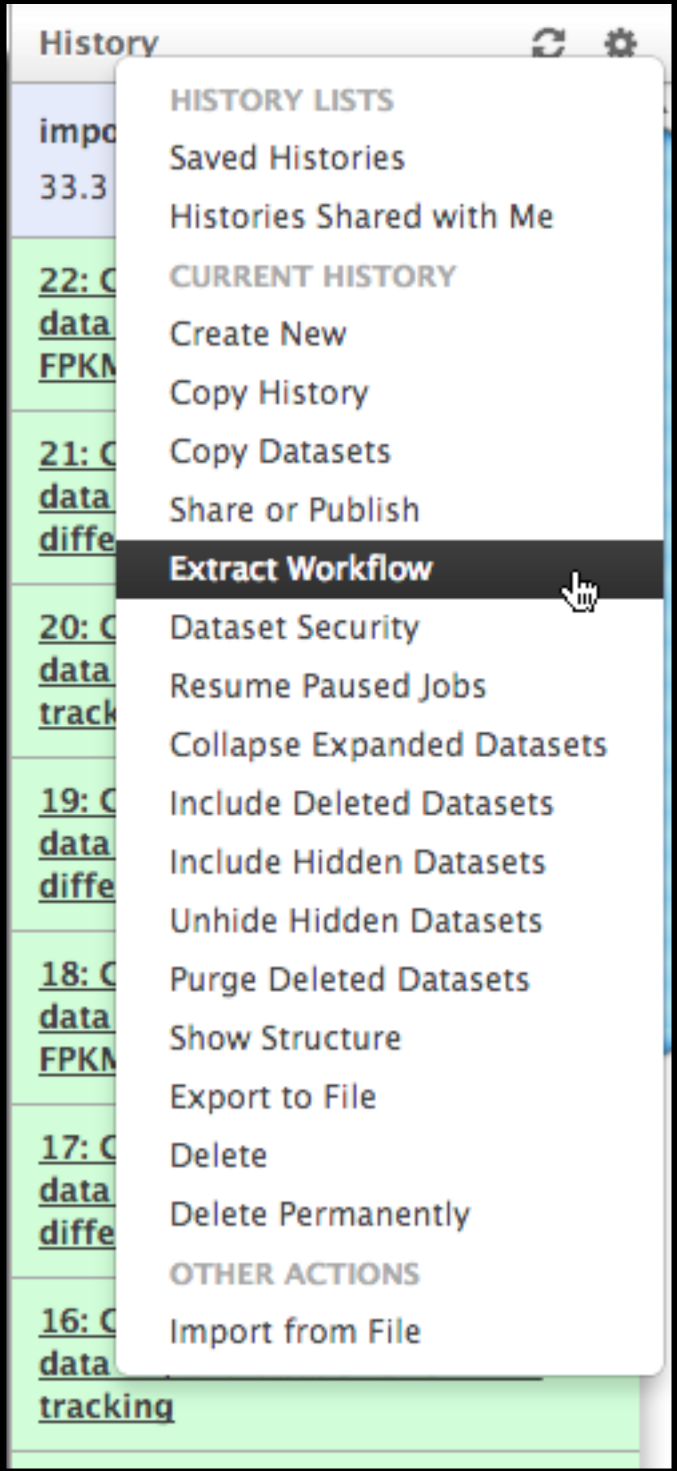
When we get in Galaxy: a generic *Overlap* Workflow

Extract Workflow from history

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

Run / test it

- Test: rerun with same inputs
- Do some on your own:
  - Count # CpG islands overlapping each exon. Did that work?
- On your own:
  - Count # of exons in each repeat
    - Did that work? *Why not?*
    - Edit workflow: doc assumptions



Agenda

Galaxy project mission

Who’s on the team

Overview & Terminology

Graphical Example - 101

Enough now ... let’s see it!

- Wiki, Tools, Histories, Sharing, Workflows, etc.

Join the Galaxy Community

Tool Shed
Mailing Lists (Sci User, Development, Private Data help)
Screencasts (Community and Galaxy Team)
Events Calendar, News Feed, Twitter, Monthly Updates
Distributions & Release Notes/Feature Descriptions
Community Wiki
Local/Public Installs, GalaxyAdmins Monthly Mtg
CiteULike group, Mendeley mirror
Annual **Galaxy Community** Meeting - GCC2013 is next!

http://galaxyproject.org/wiki/Get Involved

Registration & abstract submission are open

http://galaxyproject.org/GCC2013

Thanks for using Galaxy !!

Acknowledgments

Ann Zweig, Bob Kuhn, Jim Kent, David Haussler
& the UCSC Genome Browser Project Team and CBSE staff

Dave Clements, Anton Nekrutenko,
James Taylor, & Galaxy Team

Galaxy Community
& You!!

Areas Covered in Live Demo
1. Wiki home wiki http://wiki.galaxyproject.org
2. Search function: example -user "tophat"
3. Learn -> Datasets: custom genomes, mng datasets
4. Support: Help! Common Solutions
5. Get galaxy: asked volunteer to try/time local set-up
6. Use Galaxy -> Main: review, go to usegalaxy.org
7. Registration, Login, User & Help menu
8. UI Orientation: Tools vs History, Tool bar -> Objects
9. Pages: link live to slides - 101 tutorial, metagenomics pub, point out others
10. Using Galaxy: show all prots, import prot1, history/tool/dataset review
re-run, tool search, create workflow, sharing, edit w/ hidden & anno, run. Let run.
11. Tools: Get Data -> UCSC; Text Manipulation, Join, Sort, Unix type;
liftOver and Extract -> From UCSC; MAF tools (with paper refs); Interval Ops
12. Back to workflow (#10): review output, same as original. Unhide/hide datasets.
13. Q & A: submitted written and from audience, incorporated during talk or after