

Galaxy Project Update

2013 GMOD Meeting
Cambridge, UK

Dave Clements
Emory University



Agenda

- Project Introduction
- Project Update

What is Galaxy?

An open, web-based platform for **accessible, reproducible,**
and **transparent** computational biomedical research.

<http://galaxyproject.org>

Who here has **not** *tried* Galaxy?

```
if percentVeterans < 66:  
    demoSuccess = attemptThreeMinuteDemo()  
  
if percentVeterans >= 66 or not demoSuccess:  
    handwaveOverScreenshot()
```

<http://usegalaxy.org>

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

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

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 NSE, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

Galaxy build: \$Rev 8778:7c3df0bcbc225

galaxyproject

- intermineorg Take a look at our new interactive web services docs: iodocs.labs.intermine.org/flymine-beta 15 hours ago · reply · retweet · favorite
- galaxyproject Jackson Lab surveying bioinformatics cores. Scientific computing svy.mk/X905zC Bioinformatics and stats svy.mk/W7637u 15 hours ago · reply · retweet · favorite
- galaxyproject GCC2013 registration and abstract submission are now open bit.ly/GCC2013Reg 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/wgEncodeSydhTfbsMelCtcfDmso201qgyaleRawDataRep1.fastq>

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

Open Source Software: 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*

- 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



<http://usegalaxy.org/cloud>

Agenda

- Project Introduction
- Project Update

Software



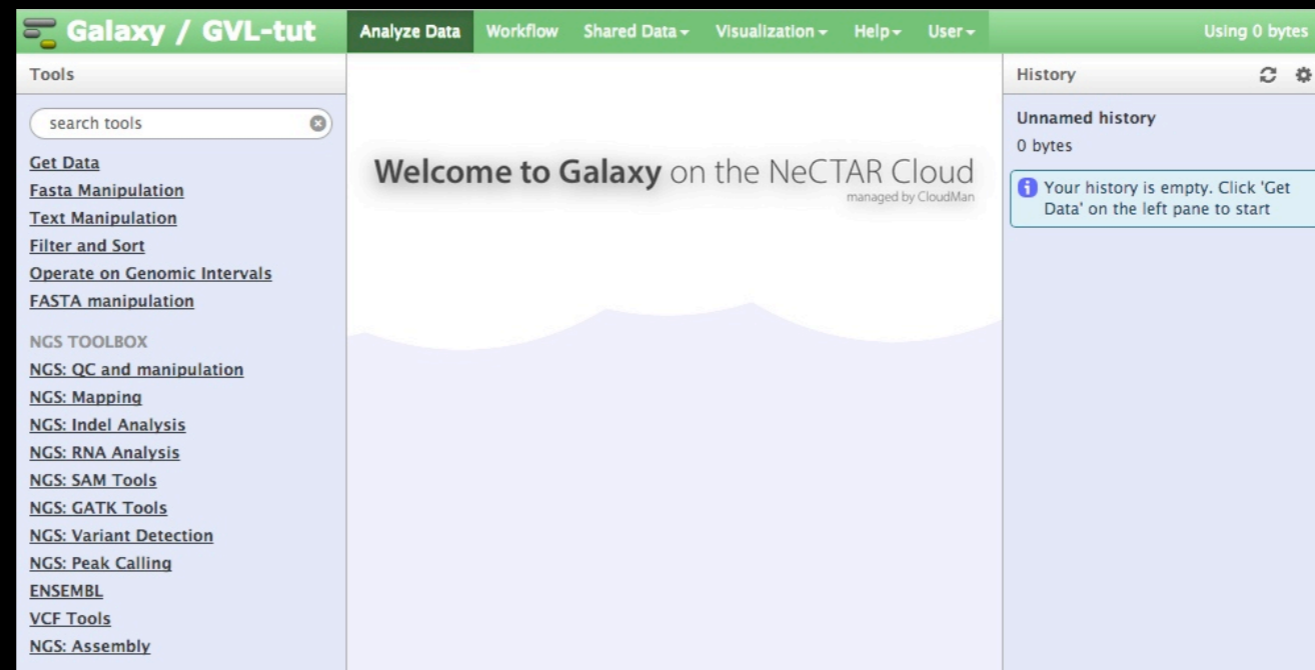
Community

Software



Community

CloudMan Platforms

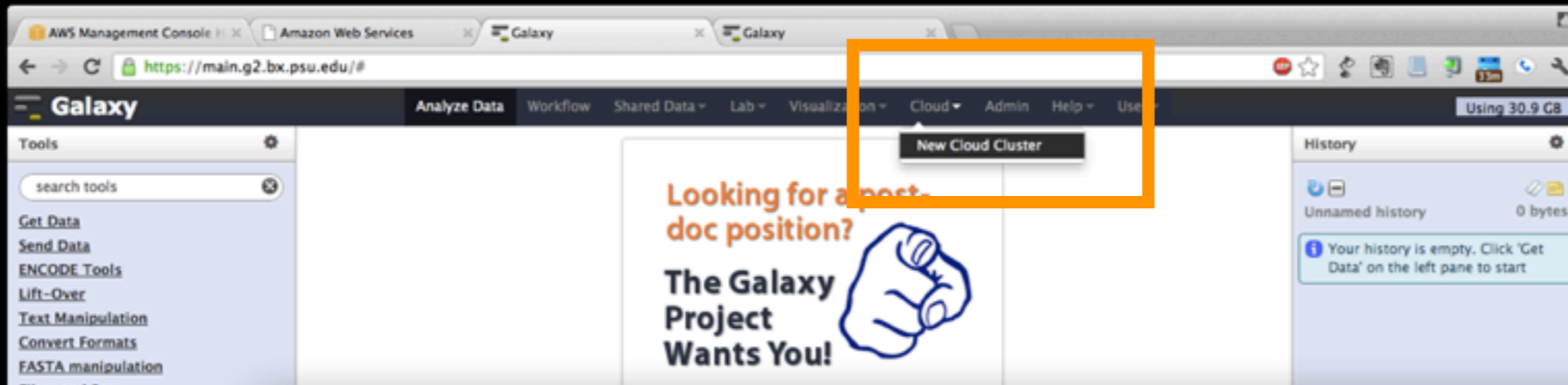


GVL:
OpenStack
NecTAR

OpenNebula:
NBIC server
Andromeda



CloudMan CloudLaunch



This screenshot shows the 'Launch a Galaxy Cloud Instance' form. The form includes the following fields and options:

- Key ID:** A text input field containing a redacted value.
- Secret Key:** A text input field containing a redacted value.
- Instances in your account:** A dropdown menu with 'New Cluster' selected.
- Cluster Name:** A text input field containing 'BTG-2012-Sept-26'.
- Cluster Password:** A text input field containing six asterisks.
- Key Pair:** A dropdown menu with 'cloudman_keypair' selected.
- Instance Type:** A dropdown menu with 'Extra Large' selected.

Below the form, there is a message: 'Requesting the instance may take a moment, please be patient. Do not refresh your browser or navigate away from the page' and a 'Submit' button.

Launch a cloud instance from another running Galaxy

Visualization: Trackster



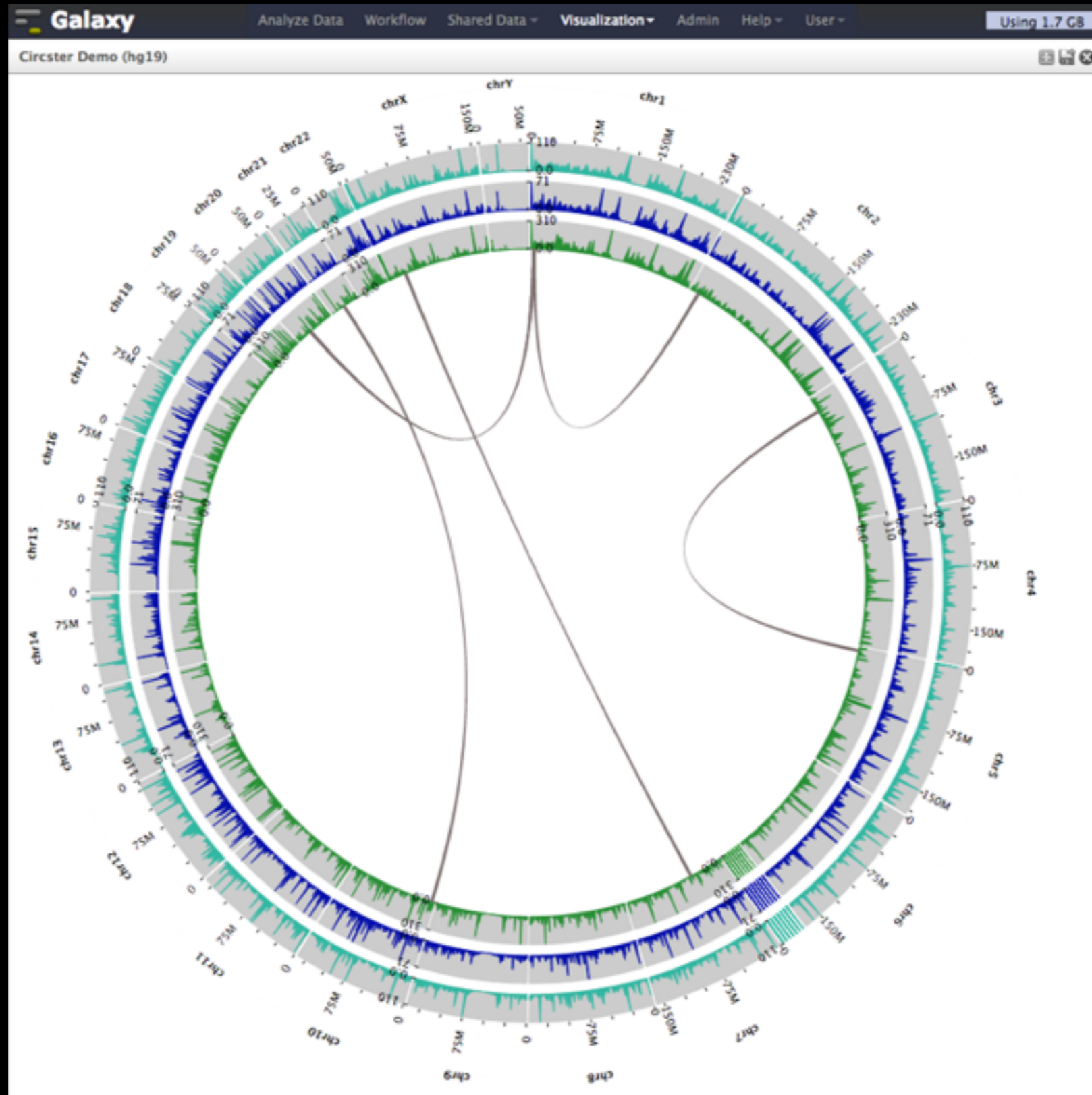
Initially Trackster; now a general purpose framework for visualization

Visualization: PhyloViz

The screenshot displays the Galaxy web interface for a Phylogenetic Tree visualization. The main window shows a tree with nodes labeled with IDs such as CED4_CAEL, 31_CAEBR, 28_DROPS, Dark_DROME, 29_AEDAE, 30_TRICA, 34_BRAFL, 35_BRAFL, 8_BRAFL, 20_NEMVE, 21_NEMVE, 9_BRAFL, 3_BRAFL, 2_BRAFL, 19_NEMVE, 37_BRAFL, 36_BRAFL, and 33_BRAFL. The interface includes a top navigation bar with 'Galaxy' and menu items like 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A status bar at the top right indicates 'Using 35.1 MB'. On the right side, there are two panels: 'Search / Edit Nodes' and 'PhyloViz Settings'. The 'Search / Edit Nodes' panel has a search dropdown set to 'Name (containing)' with 'None' entered, and a 'Search!' button. Below it are input fields for 'Name', 'Dist', and 'Annotation', and an 'Edit' checkbox. The 'PhyloViz Settings' panel has three sliders: 'Phylogenetic Spacing (px per unit)' set to 250 (range 50-1500), 'Vertical Spacing (px)' set to 18 (range 5-30), and 'Font Size (px)' set to 10 (range 5-20). 'Reset' and 'Apply' buttons are at the bottom of the settings panel.

PhyloViz from Google Summer of Code student Tomithy Too

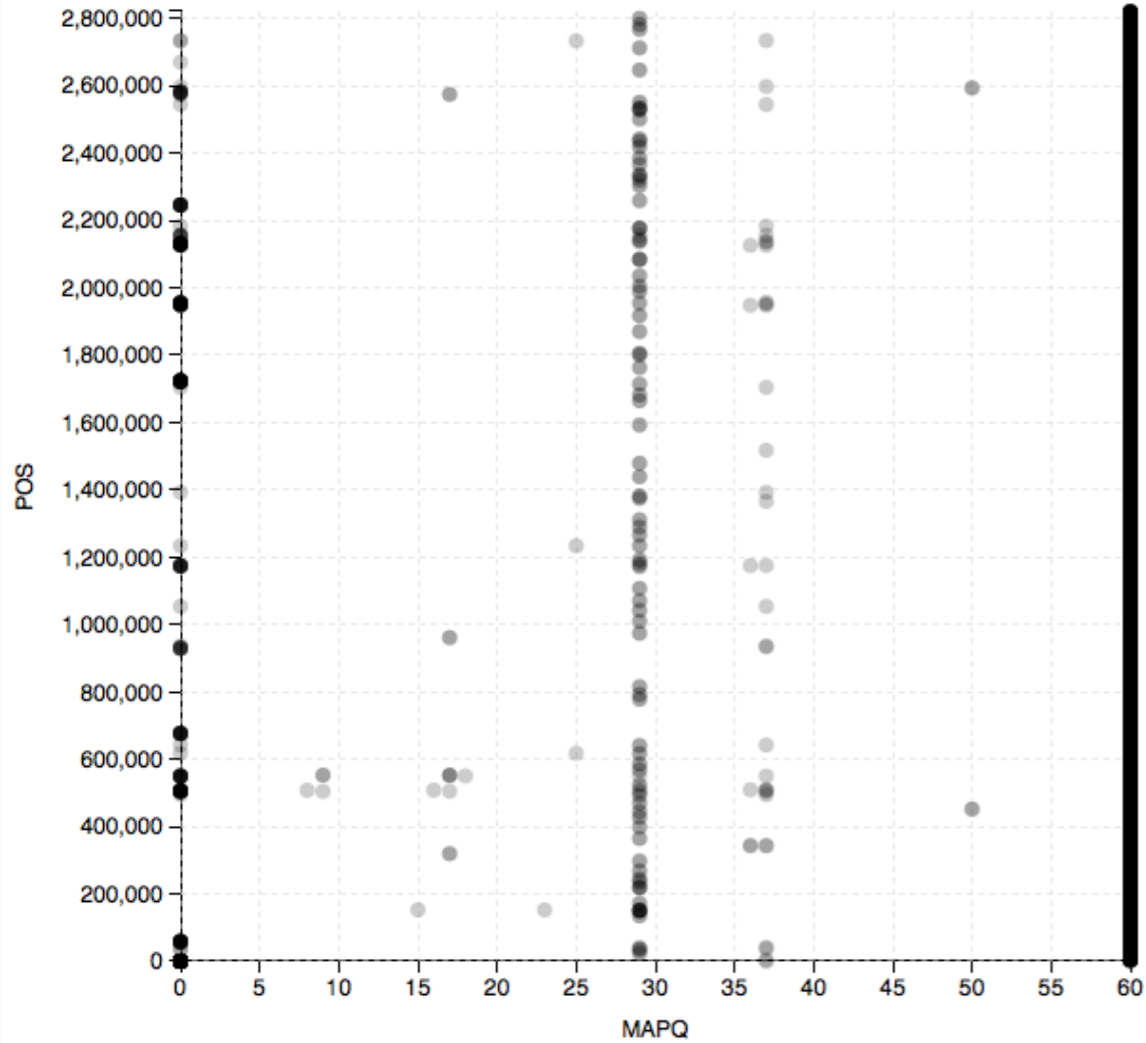
Visualization: Circster



Circster: Circos style visualizations

Visualization

Scatterplot of 'Select first on data 1'

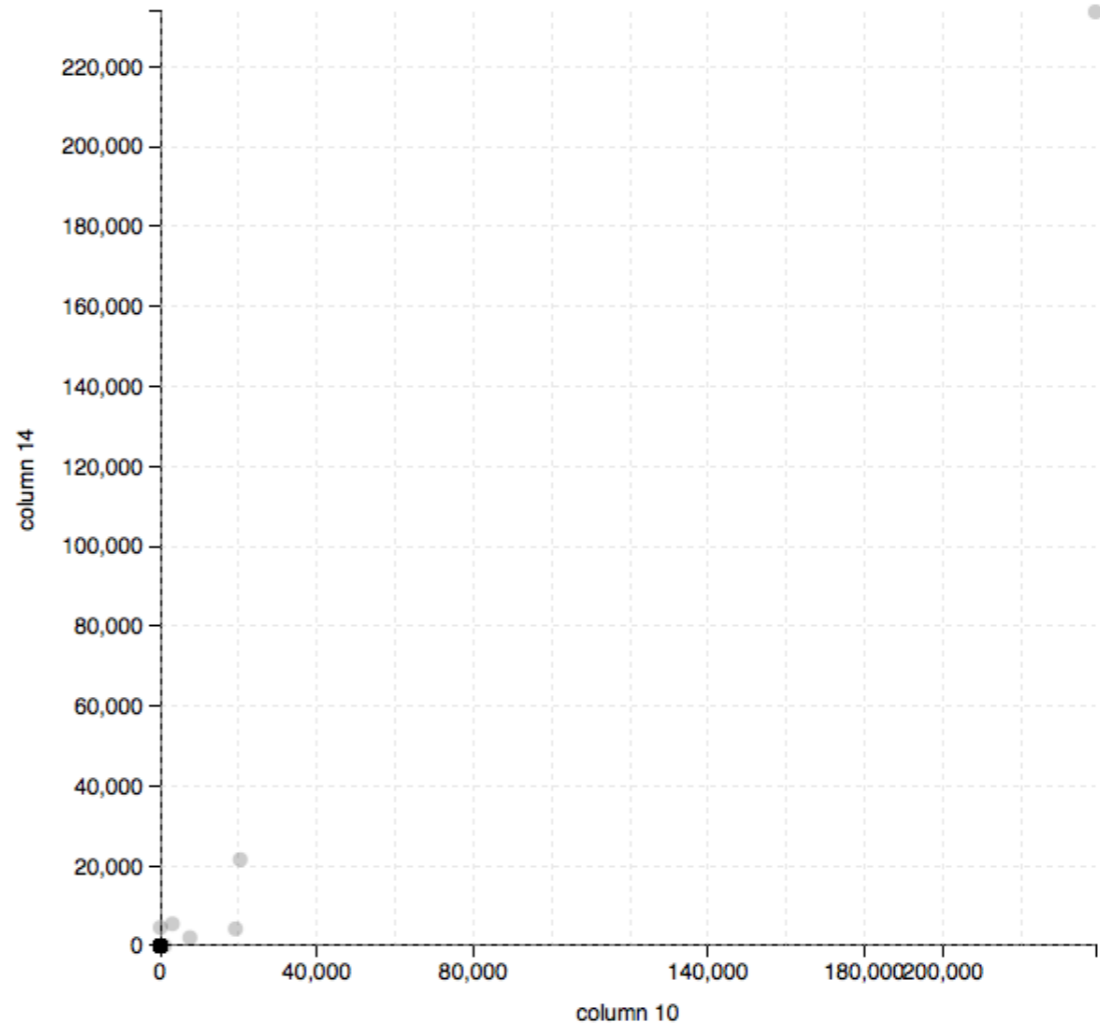


Data column for X:

Data column for Y:

Scatterplot of 'Cuffdiff on data 13, data 17, and data 26 gene FPKM tracking'

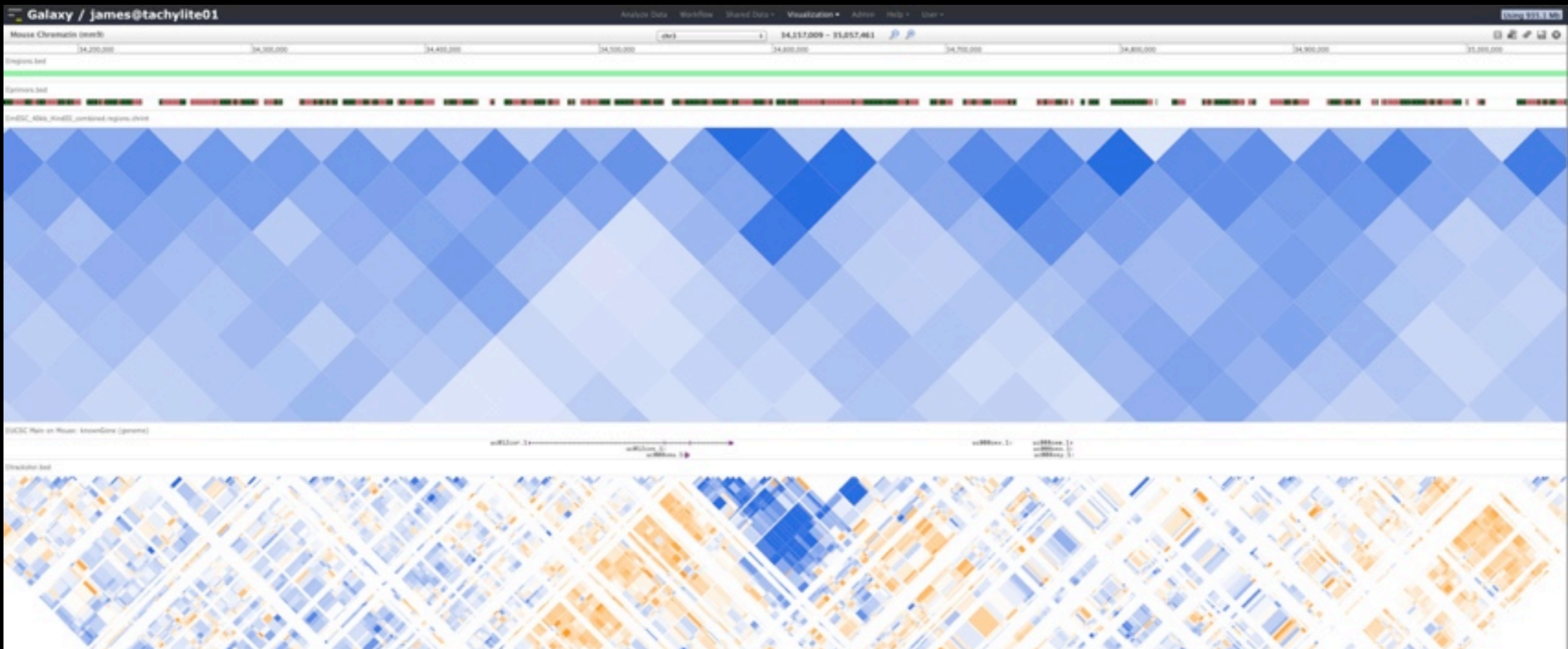
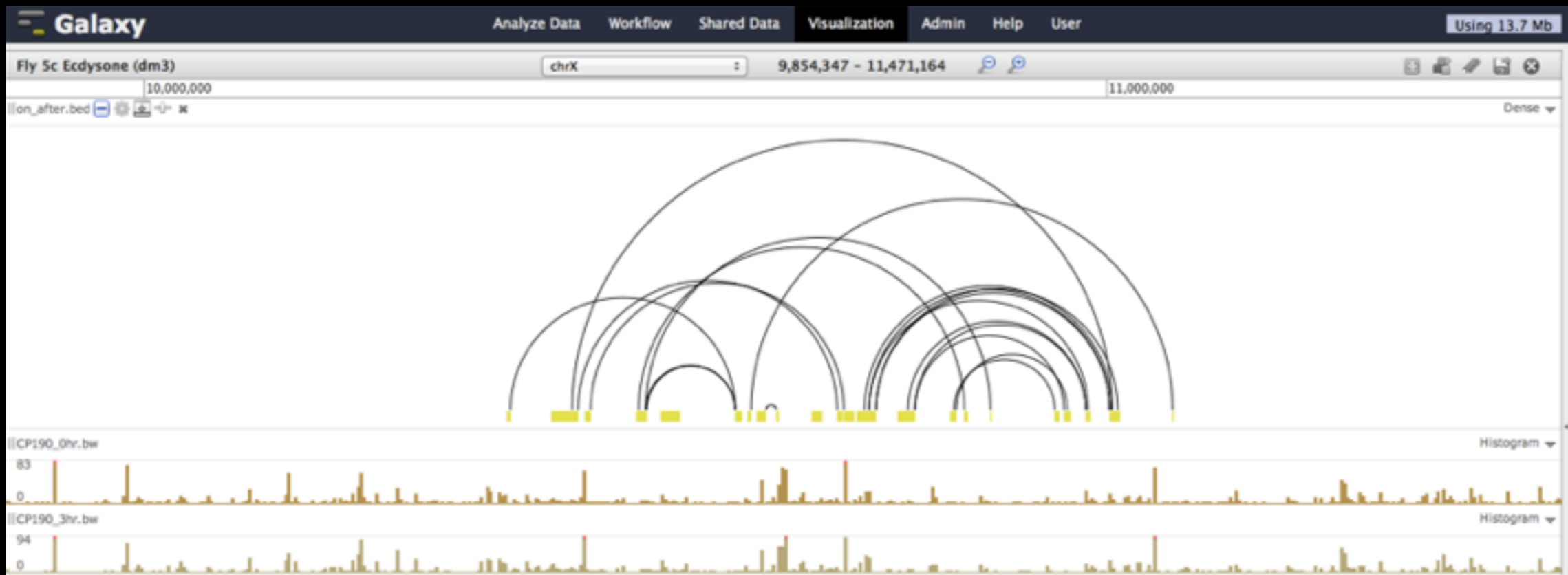
uploaded tabular file



Data column for X:

Data column for Y:

Scatter plots



Visualizations: Enhancer / Promotor Loops & Chromatin Interaction

Visual Analytics

Galaxy

Analyze Data Workflow **Shared Data** Visualization Help User

Published Visualizations | jeremy | GCC2011-2: Dynamic Filter chr19 490,747 - 2,161,375

1,000,000

GM12878 Cufflinks assembled transcripts BEST

h1-hESC Cufflinks assembled transcripts BEST

Score [68-1000]

exon_number [1-1]

frac [0-1]

cov [1-1658]

conf_lo [0-178180]

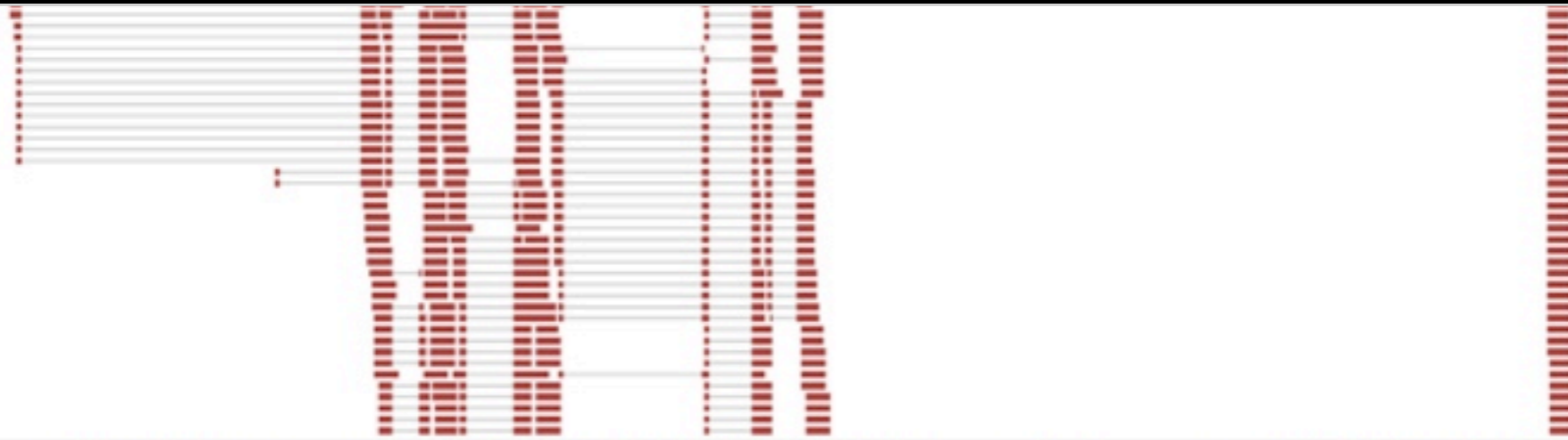
FPKM [8602-223870]

conf_hi [304-269560]

Run on complete dataset

Dynamic filtering on element properties (here, FPKM for putative transcripts)

Visual Analytics



||| h1-hESC Cufflinks assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, 0, No] - region=[all], parameters=[150000, 0.5, 0.05, 0, No] ▾

Cufflinks

Max Intron Length	<input type="text" value="300000"/>
Min Isoform Fraction	<input type="text" value="0.01"/>
Pre mRNA Fraction	<input type="text" value="0.01"/>
Min SAM Map Quality	<input type="text" value="0"/>
Perform quartile normalization	<input type="button" value="No"/>

CUFF .2427.1

→ Cufflinks - region=[chr19:1053875-1063683], parameters=[300000, 0.5, 0.05, 0, No] ▾

CUFF .2423.1

CUFF .2425.1

→ Cufflinks - region=[chr19:1053875-1063683], parameters=[300000, 0.01, 0.05, 0, No] ▾

CUFF .2435.1

CUFF .2425.1

CUFF .2427.3

→ Cufflinks - region=[chr19:1053875-1063683], parameters=[300000, 0.01, 0.01, 0, No] ▾

CUFF .2819.1
CUFF .2819.2

CUFF .2821.1

CUFF .2835.4

Modifying Cufflinks parameters and locally reassembling

Big Data: Supporting Analysis on a Massive Scale

Common request: run tools / workflows on many samples

Run each of a few dozen (paired) samples through a workflow of several dozen steps, and aggregate the results in some way

A simple analysis quickly results in dozens of workflow invocations and hundred of individual tool runs

Big Data: Plans

Rewrite default workflow engine

Histories will be able to contain pending workflows, dataset groups, other entities - not just datasets

Rather than scheduling all at once, monitor workflow progress, allow pausing in response to failure or user intervention, decision nodes, streaming data and intermediate datasets, ...

Make workflow scheduling engine pluggable

Once it is a background process, can afford the time to delegate

Pluggability / Extensibility / APIs

- Workflow rewrite
- Visualization framework
- ObjectStore storage api
- Galaxy API
- ...
- Make everything pluggable; start using those interfaces internally.

Software



Community

Software



Community

Release Cycle

Experimented with 2-3 week release cycle

Now settled on 2 month release cycle

Less thrashing for us and users

Better testing and doc



Galaxy Code documentation » lib » galaxy Package » webapps Package » galaxy Package » [previous](#) [next](#) [modules](#) [index](#)

Galaxy API Documentation

Background

In addition to being accessible through a web interface, Galaxy can now also be accessed programmatically, through shell scripts and other programs. The web interface is appropriate for things like exploratory analysis, visualization, construction of workflows, and rerunning workflows on new datasets.

The web interface is less suitable for things like

- Connecting a Galaxy instance directly to your sequencer and running workflows whenever data is ready
- Running a workflow against multiple datasets (which can be done with the web interface, but is tedious)
- When the analysis involves complex control, such as looping and branching.

The Galaxy API addresses these and other situations by exposing Galaxy internals through an additional interface, known as an Application Programming Interface, or API.

Quickstart

Log in as your user, navigate to the API Keys page in the User menu, and generate a new API key. Make a note of the API key, and then pull up a terminal. Now we'll use the `display.py` script in your `galaxy/scripts/api` directory for a short example:

```
% ./display.py my_key http://localhost:4096/api/histories
Collection Members
-----
#1: /api/histories/8c49be448cfe29bc
   name: Unnamed history
   id: 8c49be448cfe29bc
#2: /api/histories/33b43b4e7093c91f
   name: output test
   id: 33b43b4e7093c91f
```

Project Versions

latest

RTD Search

Full-text doc search.

Table Of Contents

- Galaxy API Documentation
 - Background
 - Quickstart
 - API Controllers
 - datasets Module
 - folder_contents Modul
 - folders Module
 - forms Module
 - genomes Module
 - group_roles Module
 - group_users Module
 - groups Module
 - histories Module
 - history_contents Modu
 - item_tags Module
 - libraries Module
 - library_contents Modu
 - permissions Module
 - quotas Module
 - request_types Module
 - requests Module
 - roles Module

Galaxy toolshed vision

- Allow users to share “suites” containing tools, datatypes, workflows, sample data, and automated installation scripts for tool dependencies
- Version controlled
- Community annotation, rating, comments, review
- Dependency resolution
- Integration with Galaxy instances to automate tool installation and updates
- A key to intergalactic unification
- Lots and lots of progress in past 12 months

Trello

The screenshot shows a Trello board interface for 'Galaxy: Development Inbox'. The board is organized into four main columns: 'Inbox', 'Developer ideas', 'Bug Reports', and 'Issues from Bitbucket'. Each column contains several cards representing tasks or issues. The 'Inbox' column has five cards, including one about adding cards and another about a filter and sort tool. The 'Developer ideas' column has five cards, such as 'Anonymous use of workflows/visualizations' and 'Feature Request: the ability to restart a failed workflow'. The 'Bug Reports' column has six cards, including 'Issues with workflow step hiding not persisting' and 'Unable to run jobs when user job limits are set'. The 'Issues from Bitbucket' column has six cards, such as '5: Option to disable automatic history creation' and '8: More flexible output handlers'. On the right side, there is a 'Members' section with a grid of member avatars and an 'Add Members...' button. Below that is a 'Board' section with 'Options', 'Add List', and 'Filter Cards' buttons. At the bottom right is an 'Activity' section showing recent actions, such as 'Dannon Baker added API: Library Contents to Developer ideas and...' and 'g2roboto on Feature request: manually hide datasets'.

Trello | Search | Help | Notifications | Boards

Galaxy: Development Inbox | Galaxy Project | Public

Inbox

- To add cards, use the <http://galaxyproject.org/trello>
2 votes | 1 comment
- Filter and Sort: "Select" tool not dealing with special characters right
1 comment
- Uploaded fastq file datatype not usable in BWA
1 comment
- Reference genome request: GATK-ordered hg19
1 comment
- Feature request: manually hide datasets
1 comment
- Add a card...

Developer ideas

- Anonymous use of workflows/visualizations
0/2
- Feature Request: the ability to restart a failed workflow from the point of failure;
6 votes | 2 comments
- Google Drive / Dropbox / Box / ... integration
1 vote
- Bug report: always import deleted datasets
2 comments
- Standalone web application(s) for visualizations
- Enh: Archiving histories
1 comment
- Modify data library upload completion message
1 comment
- Display in UI runtime
- Add a card...

Bug Reports

- Issues with workflow step hiding not persisting
1 vote | 1 comment
- Workflow View Broken in Toolshed?
1 comment
- Unable to run jobs when user job limits are set
1 vote | 4 comments
- Fix tool tip FASTQ Summary Statistics
1 comment
- Bug when using data_column
1 comment
- Velvet wrapper broken when real user jobs are used
1 comment
- apport.fileutils
1 comment
- Bug: Running functional tests for migrated or installed tools does not
1 comment
- Add a card...

Issues from Bitbucket

- 5: Option to disable automatic history creation
2 votes | 1 comment
- 6: Option to require that histories have names
1 vote
- 8: More flexible output handlers
1 comment
- 10: Allow overriding parameters when running a workflow
1 vote
- 20: Suggestion: new tag in tool's XML file - 12/9/08 email from Assaf Gordon
1 comment
- 21: Real DB key build ontology
1 comment
- 24: Add ability to password secure tools
1 comment
- Add a card...

Members

Grid of member avatars: CE, DB, G

Add Members...

Board

Options

Add List

Filter Cards

Activity

Dannon Baker added API: Library Contents to Developer ideas and
• sent to the board
• joined
today at 10:39 am

G **g2roboto** on Feature request: manually hide datasets
Submitted by @nickstoler
Feb 1 at 4:40 pm

G **g2roboto** added Feature request: manually hide datasets to Inbox.
Feb 1 at 4:40 pm

G **g2roboto** on Reference

Software



Community

Software



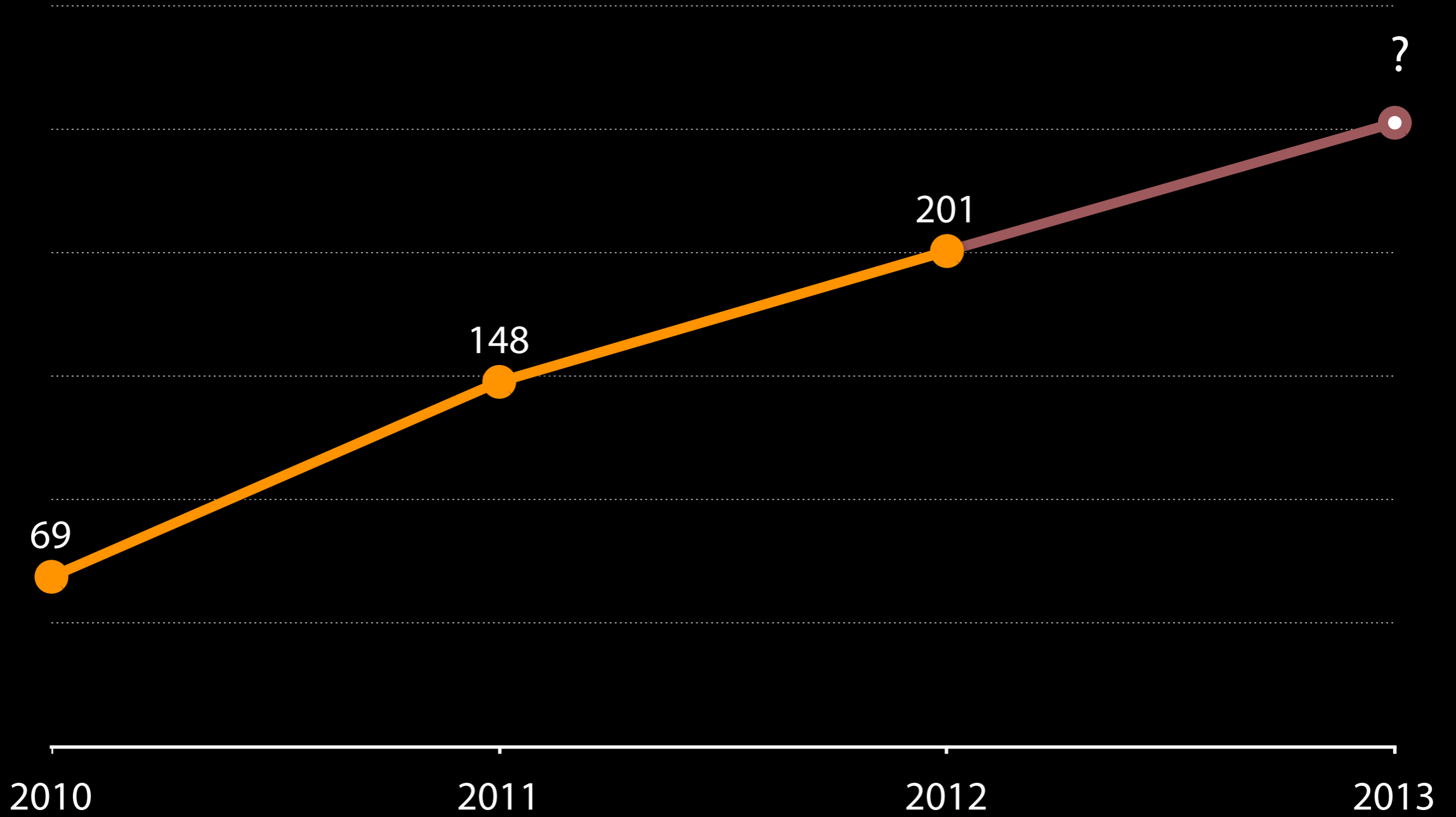
Community

2012 and 2013 Meetings



GCC2013 Registration & abstract submission open
<http://galaxyproject.org/GCC2013>

GCC attendance over time



New Communities

GalaxyAdmins

Administrators of large Galaxy Instances
Started by Ulowa in 2012

Galaxy-France

French language and France-centric Galaxy community mailing list
Launched after Galaxy Tour de France in 2012

Galaxy-Public-Servers

Mailing list for those hosting public Galaxy servers
Just launched

Mailing Lists

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

Galaxy-Announce

Project announcements, low volume, moderated

Low volume (2012: 42 posts, 1600 members)

Galaxy-User

Questions about using Galaxy and usegalaxy.org

High volume (2012: 2900 posts, 2700 members)





Galaxy-Dev

Questions about developing for and deploying Galaxy

High volume (2012: 4500 posts, 850 members)

Training

Workshops offered by Galaxy Team in 2012

January	February	March	April
 <p>2 Events 245 People 717 Participant hrs Czech Rep, CA</p> 	<p>0 Events People Participant hrs</p>	<p>1 Event 50 People 200 Participant hrs France</p> 	 <p>7 Events 225 People 1125 Participant hrs DC, MD, IA</p>  
May	June	July	August
 <p>6 Events 291 People 882 Participant hrs France, NC</p> 	 <p>3 Events 274 People 822 Participant hrs France</p>	 <p>2 Events 230 People 1330 Participant hrs France, IL (GCC)</p> 	 <p>1 Events 20 People 80 Participant hrs NC</p>
September	October	November	December
 <p>2 Events 45 People 585 Participant hrs South Africa</p>	 <p>4 Events 102 People 449 Participant hrs IL, IN</p> 	 <p>3 Events 440 People 720 Participant hrs CA</p>	 <p>1 Event 50 People 750 Participant hrs PR</p>

2012

29 Training Events
17 Universities
7 Meetings
4 Countries
8 States
3 Continents
1,677 People
6,193 Participant hours

All workshops **hands-on**. Almost all of these used **CloudMan** based servers.

Almost all supported by an **AWS in Education Grant** for Galaxy Training



Plus at *least* 13 other seminars / talks by Galaxy Team members, and talks and workshops by community members, and ...

just too much stuff to count:

<http://wiki.galaxyproject.org/Events/Archive>

Acknowledgements

GMOD:

Scott Cain

Amelia Ireland

The Galaxy Team



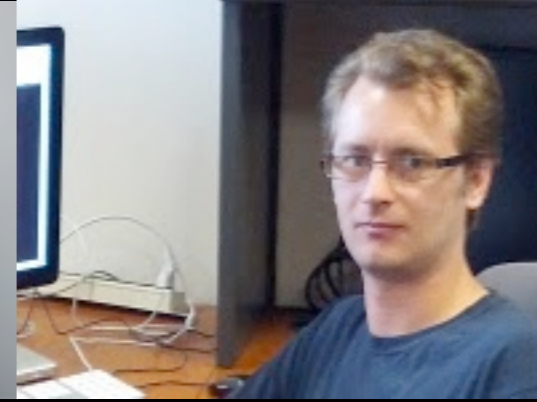
Enis Afgan



Dannon Baker



Dan Blankenberg



Dave Bouvier



Dave Clements



Nate Coraor



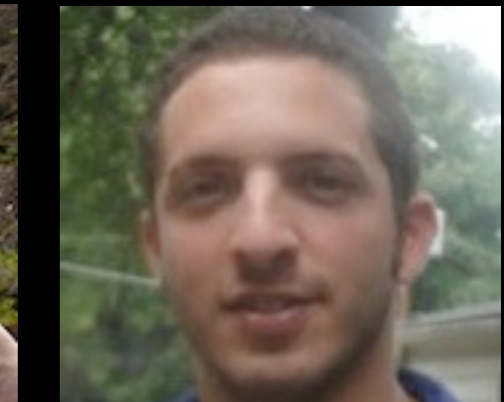
Carl Eberhard



Dorine Francheteau



Jeremy Goecks



Sam Guerler



Jen Jackson



Greg von Kuster



Ross Lazarus



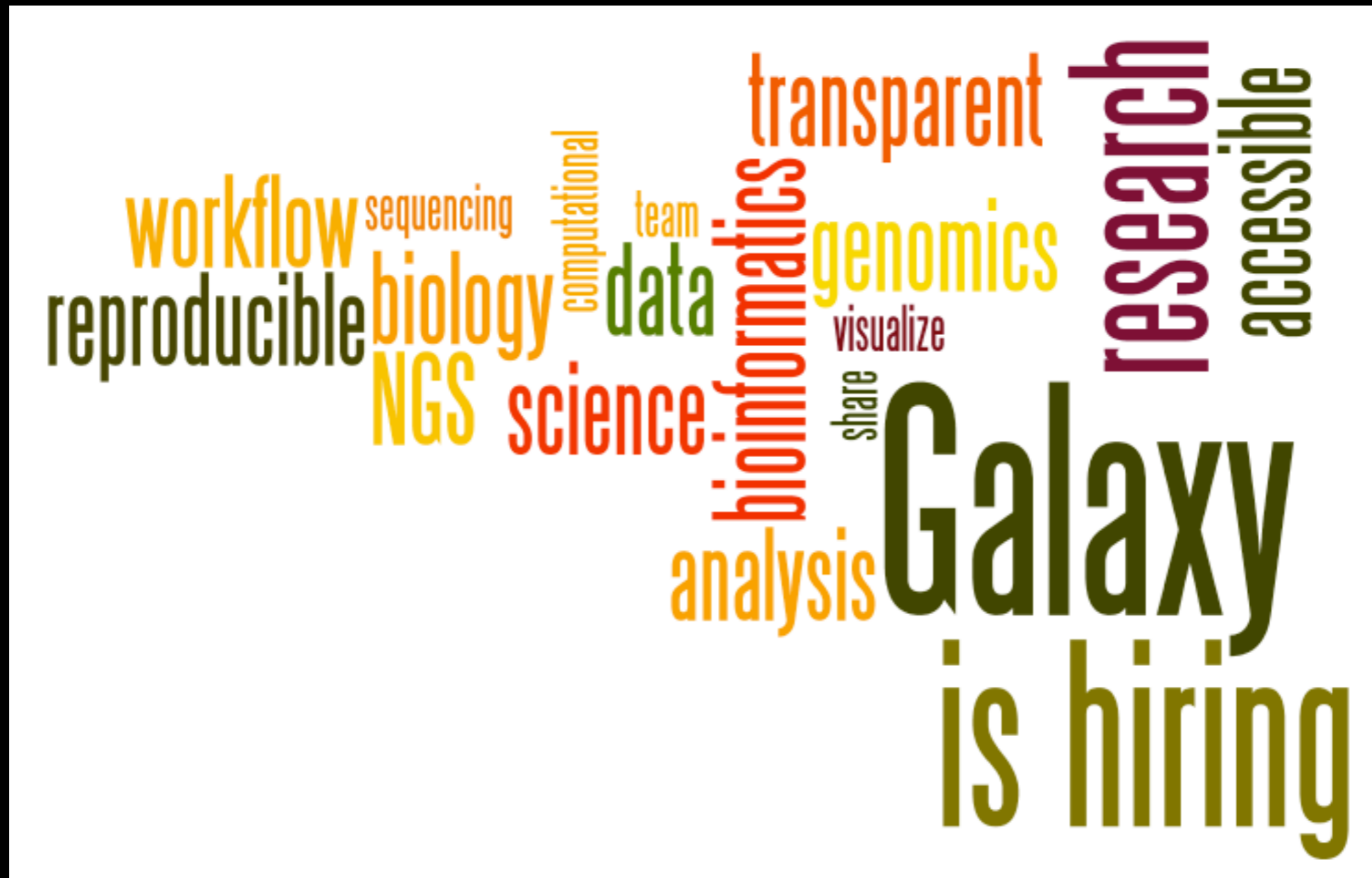
Anton Nekrutenko



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>



Thank You!