Galaxy Project Update

2013 GMOD Meeting
Cambridge, UK

Dave Clements
Emory University

Galaxy

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

handWaveOverScreenshot()

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
  - Intersect the intervals of two datasets
  - Subtract the intervals of two datasets
  - Merge the overlapping intervals of a dataset
  - Concatenate two datasets into one dataset
  - Base Coverage of all intervals

Join (version 1.0.0)

Join:
1: Exons
First dataset

with:
2: Repeats
Second dataset

with min overlap:
1
(bp)

Return:
Only records that are joined (INNER JOIN)

Execute

TIP: If your dataset does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.

Screencasts!

See Galaxy Interval Operation Screencasts (right click to open this link in another window).

Syntax

Where overlap specifies the minimum overlap between intervals that allows them to be joined.
Return only records that are joined returns only the records of the

History

Basic Protocol 4
10.4 MB

13: Join on data 2 and data 1
267 regions
format: interval, database: hg19
display at UCSC main
view in GeneTrack
display at Ensembl Current
display at RViewer main

1.Chrom	2.Start	3.End	4.Name
chr22	17586742	17586844	NM_01433
chr22	17600280	17602017	NM_03189
chr22	17600280	17602017	NM_03189
chr22	17600280	17602017	NM_03189
chr22	17600280	17602017	NM_03189
chr22	17600280	17602017	NM_03189
chr22	17600280	17602017	NM_03189

12: Cluster on data 2
11: Complement on data 2
10: Coverage on data 2 and data 1
9: Base Coverage on data

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

NIOO nbic SURF SARA BiG Grid

Andromeda: A cloud-based Galaxy

Live Quickies

ic fastQ ipulation: | Advanced fastQ manipulation: | 454 Mapping: Single End | Uploading Data using FTP | Managing account histories

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: $Rev 8778:7c3df0bcbc22$

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/X9O5zC 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 dataaset for CPB Chip-Seq prorocol
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/wgEncodeSydhTfbsMelCtcfDmso20IggyaleRawDataRep1.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

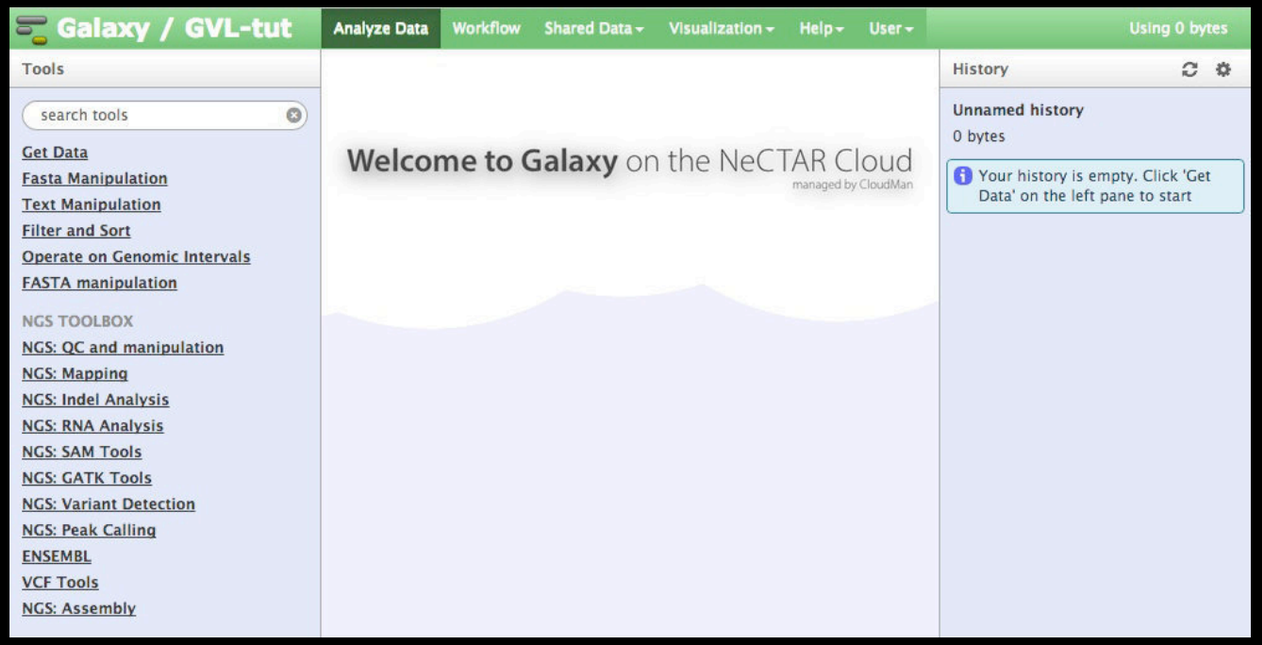
Community

Software

Community

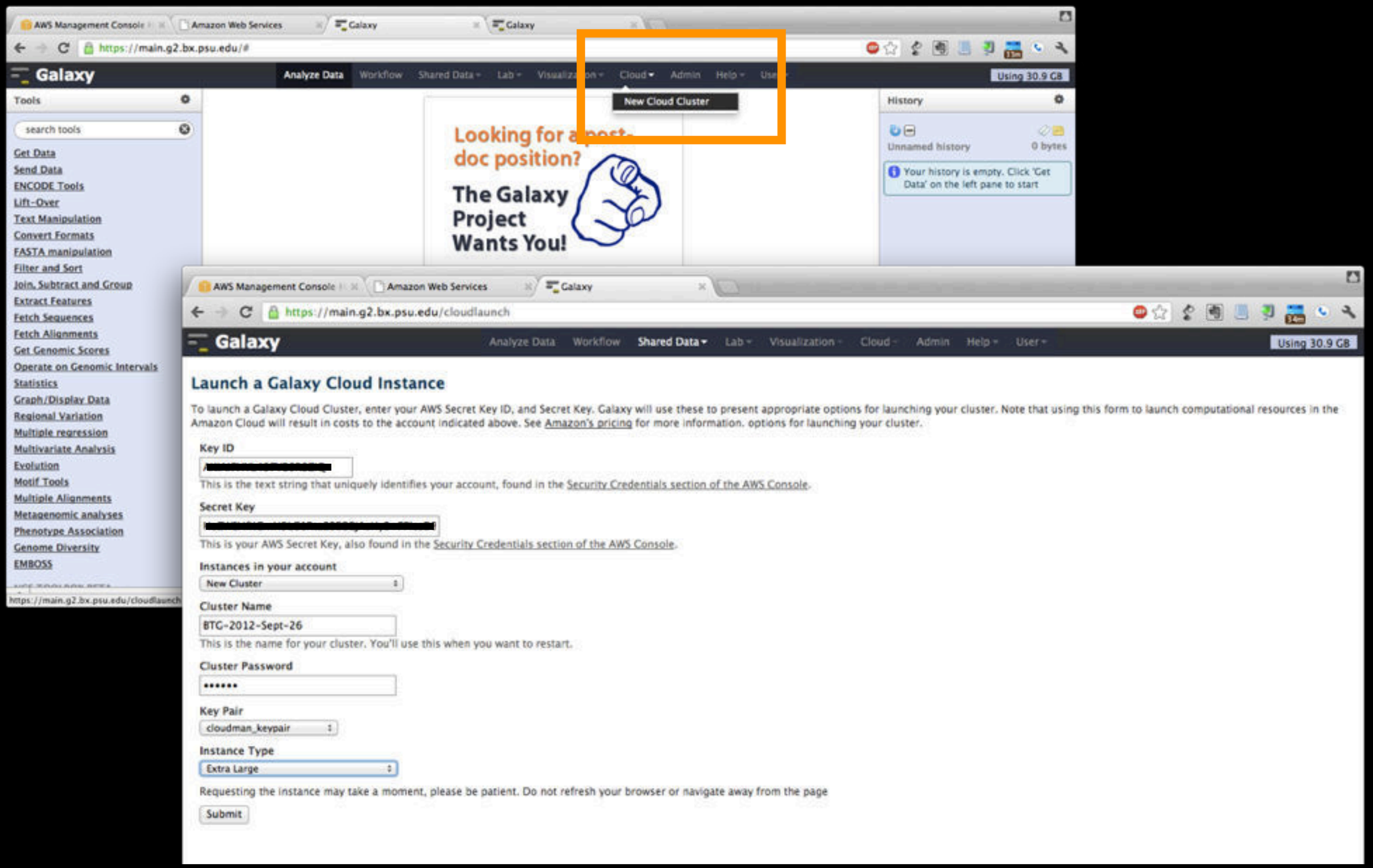
CloudMan Platforms

GVL:
OpenStack
NecTAR



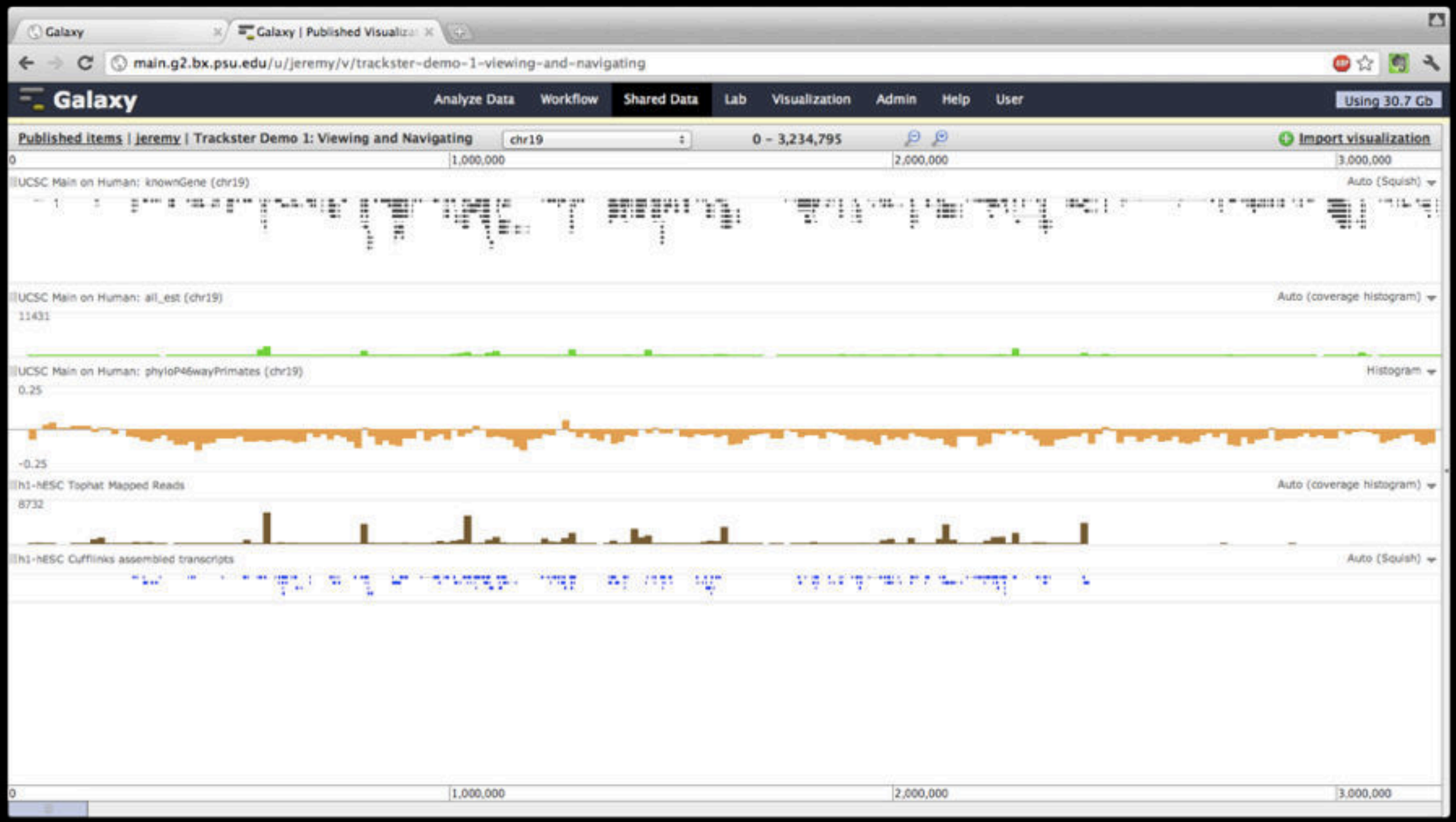
OpenNebula:
NBIC server
Andromeda

CloudMan CloudLaunch



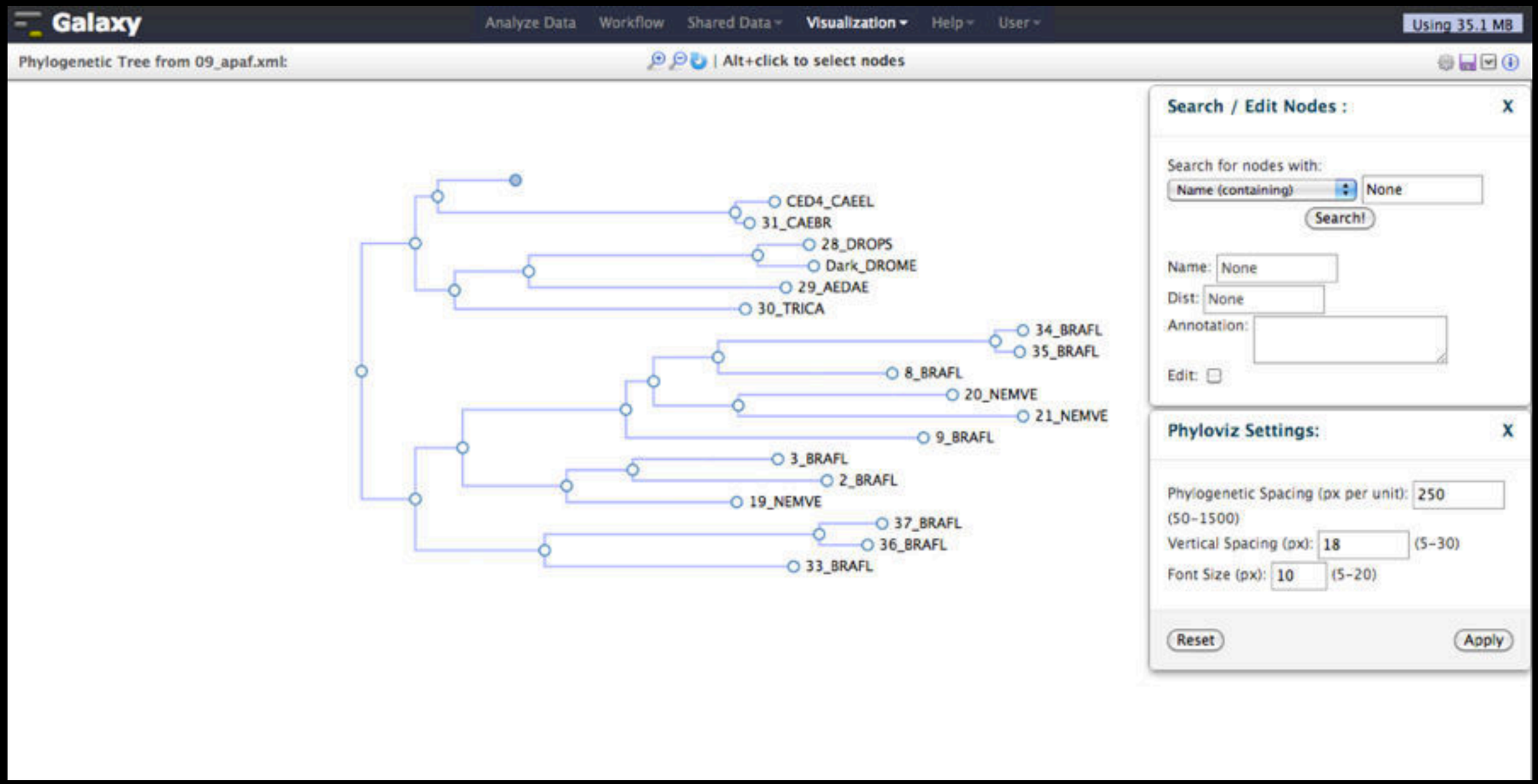
Launch a cloud instance from another running Galaxy

Visualizatiion: Trackster



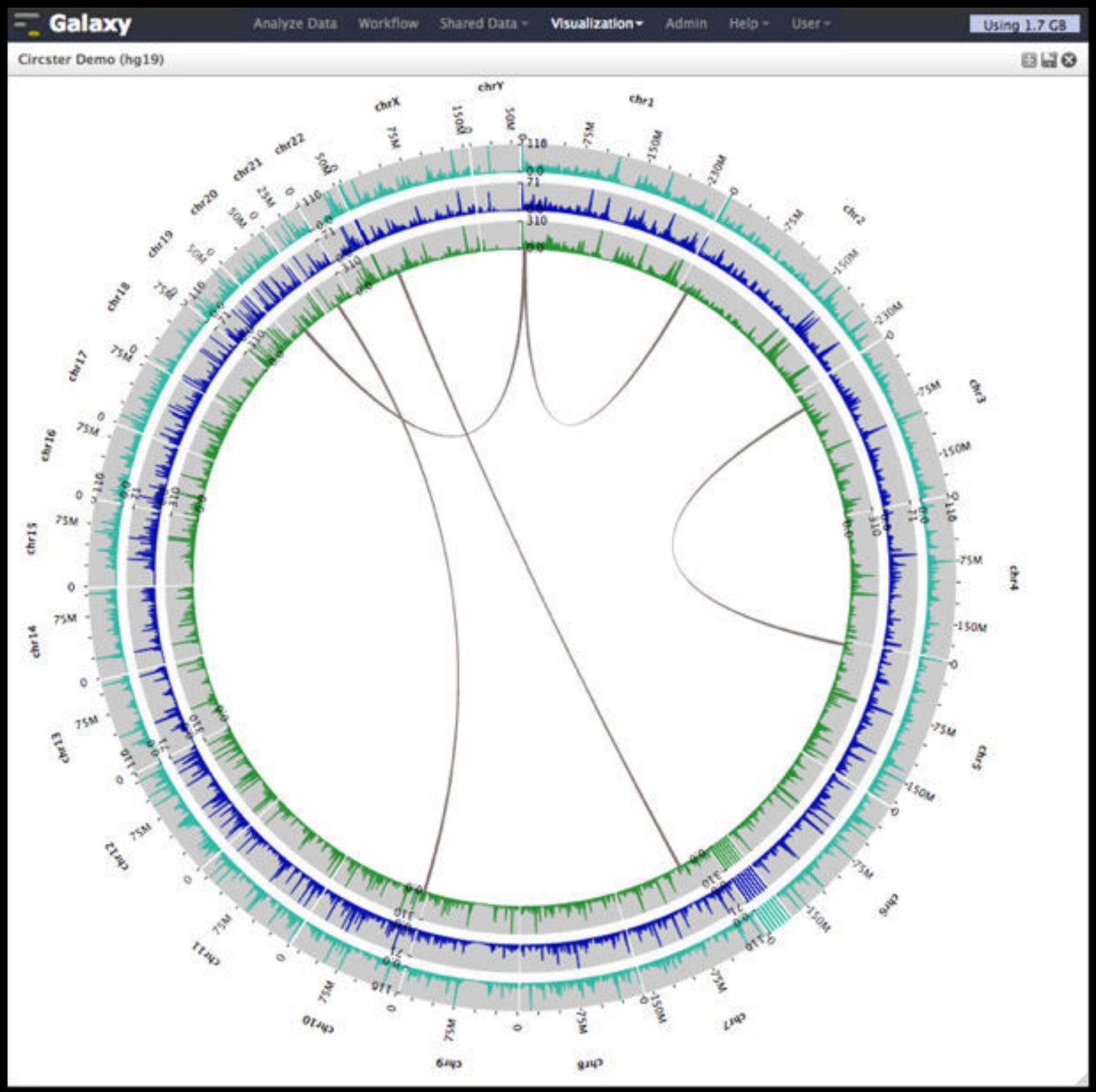
Initially Trackster; now a general purpose framework for visualization

Visualization: PhyloViz



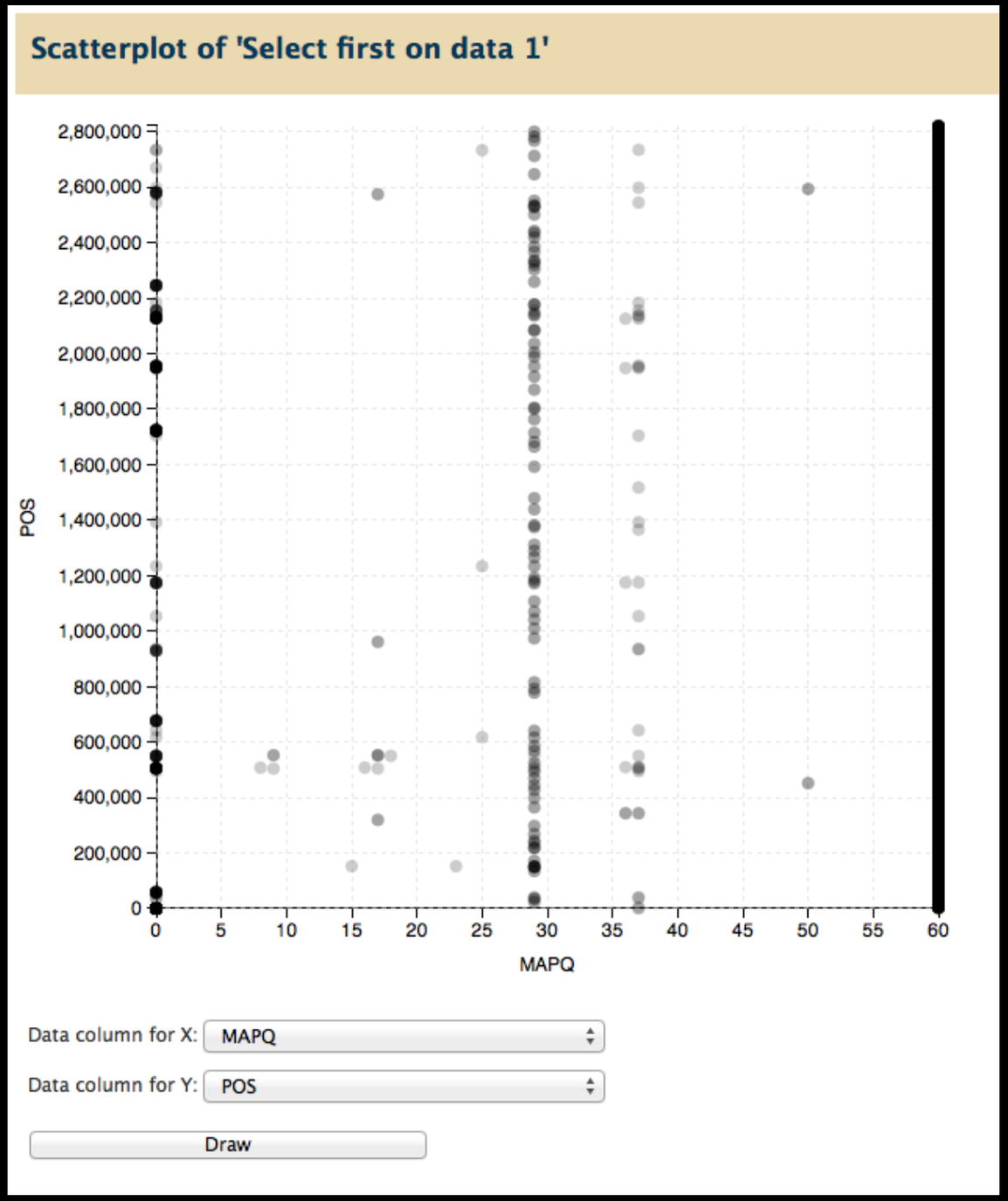
PhyloViz from Google Summer of Code student Tomithy Too

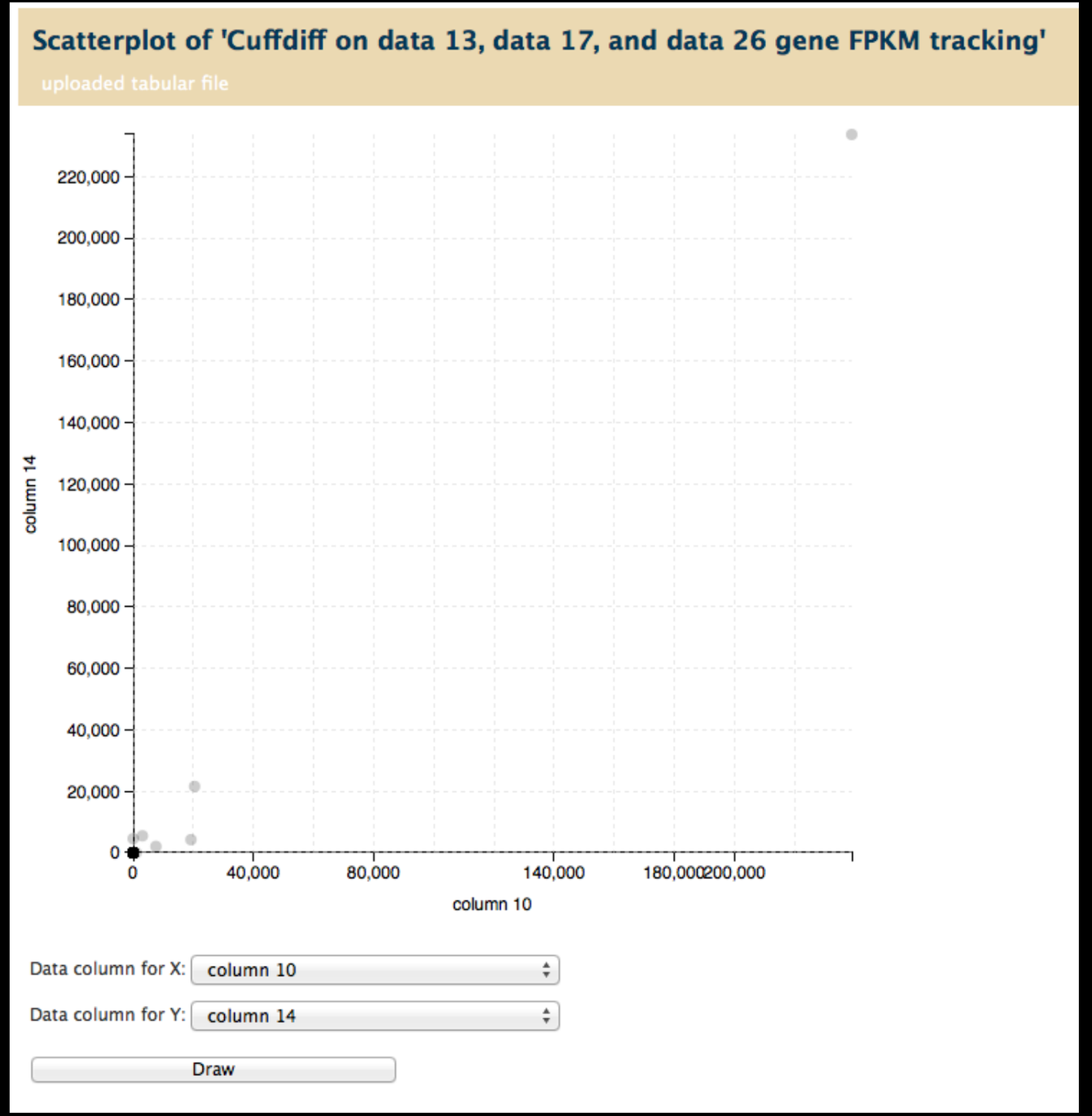
Visualization: Circster



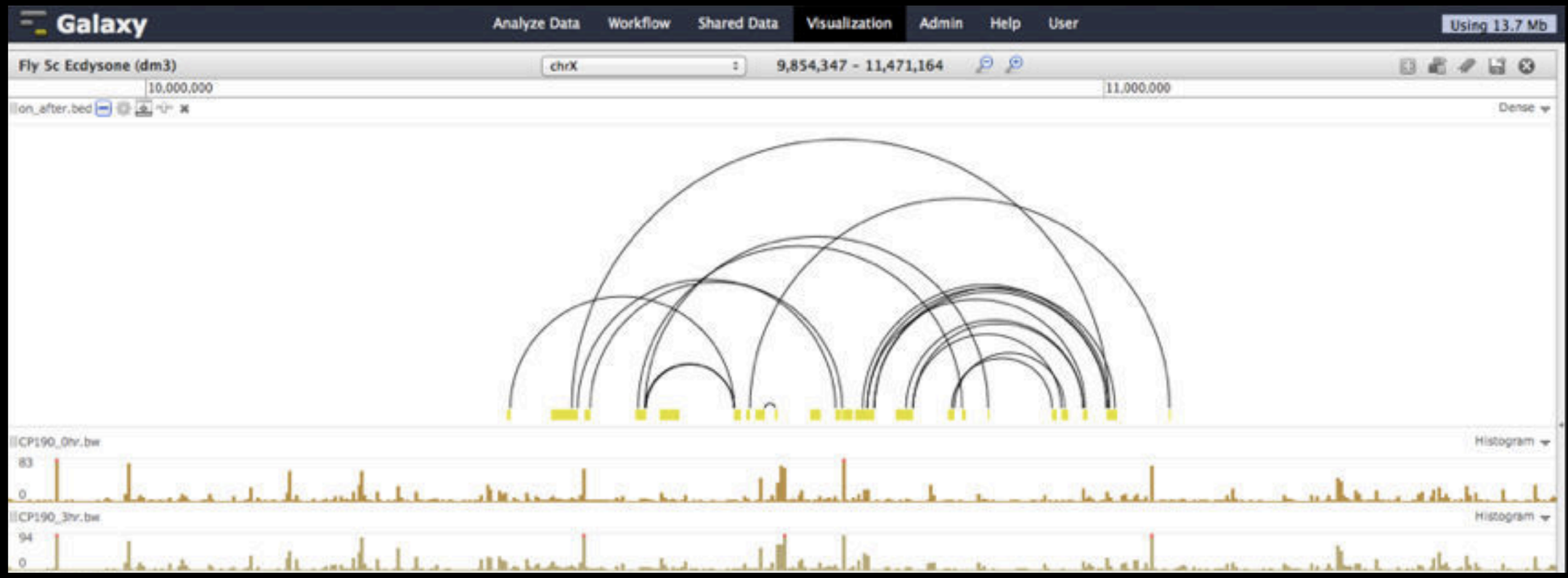
Circster: Circos style visualizations

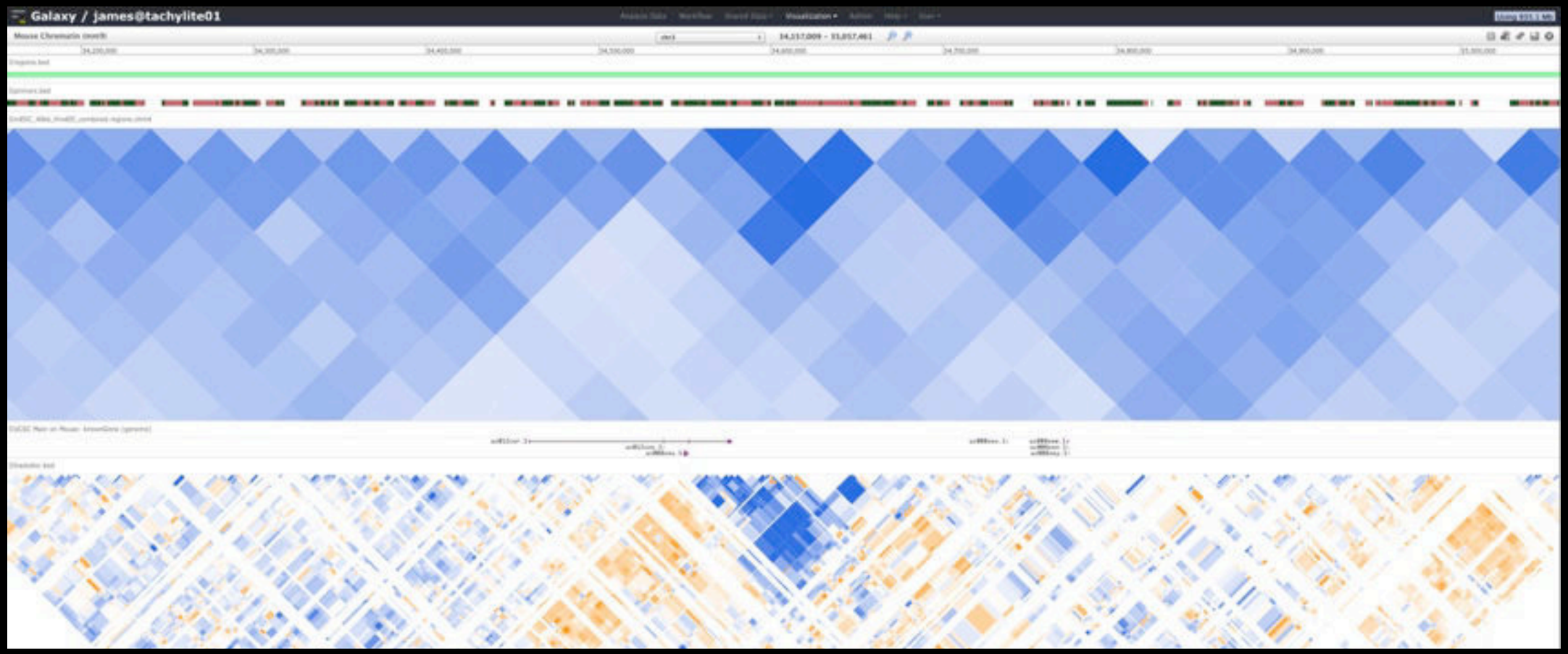
Visualization





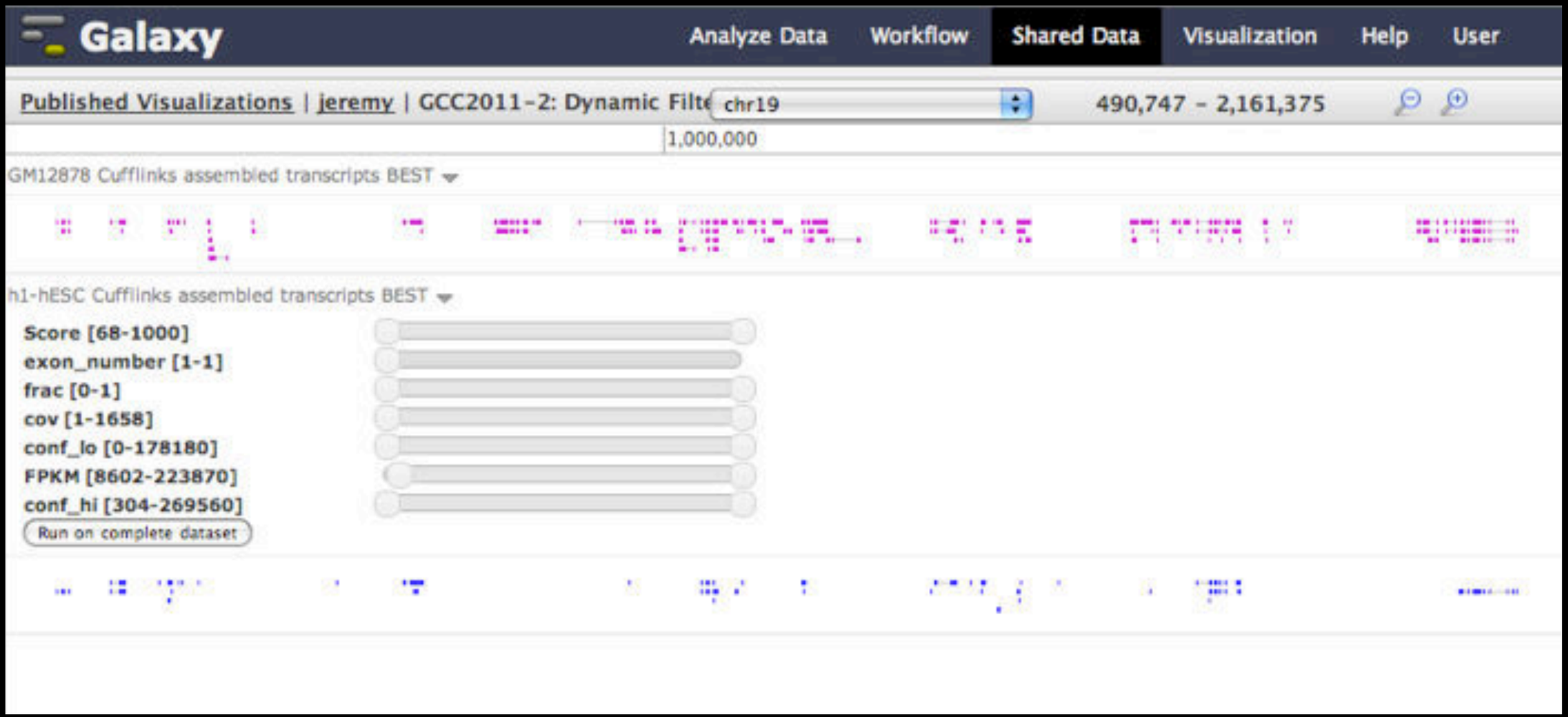
Scatter plots





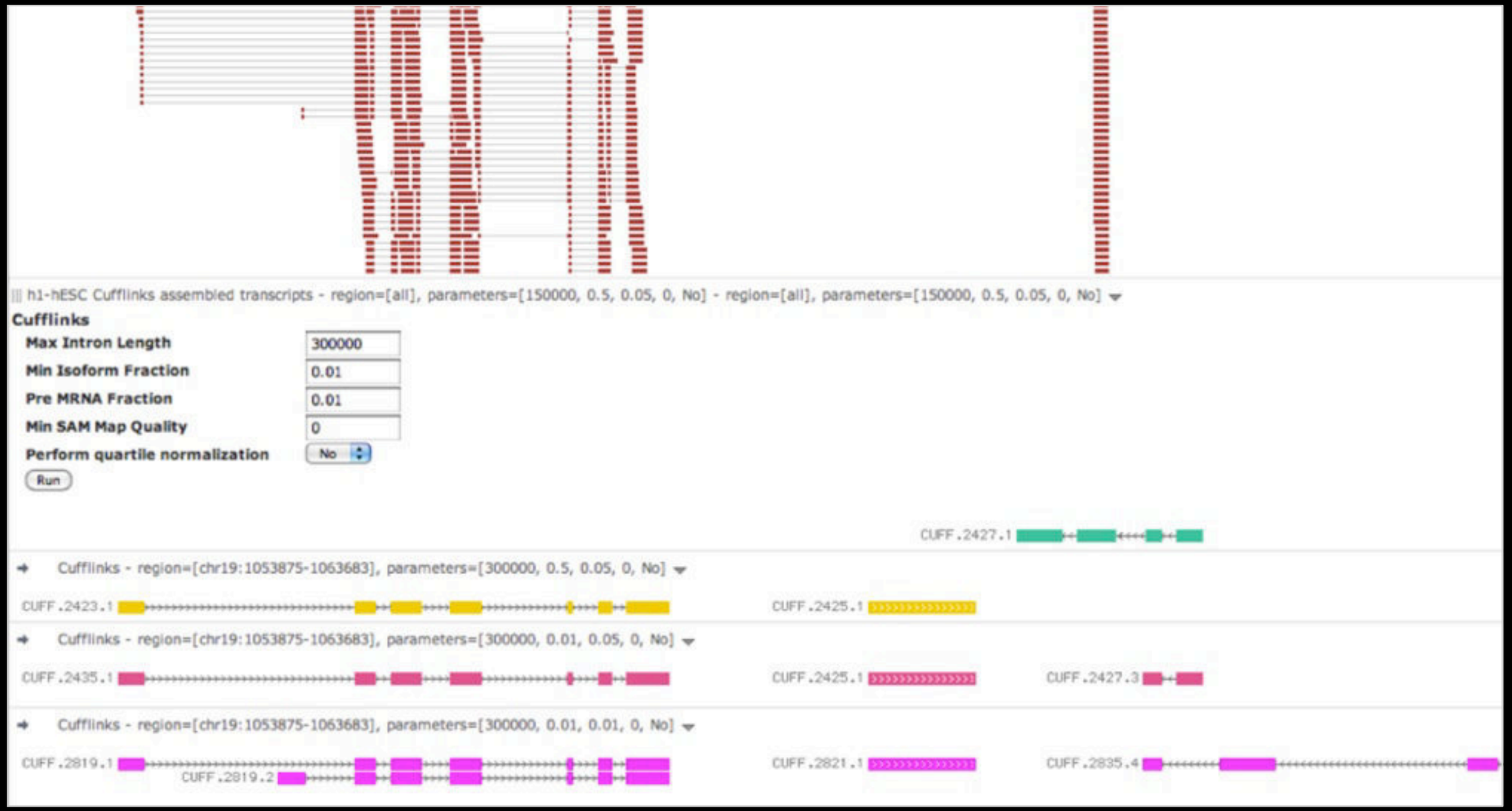
Visualizations: Enhancer / Promotor Loops & Chromatin Interaction

Visual Analytics



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

Visual Analytics



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
- Vizualization 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

Code Doc

Galaxy Code documentation » lib » galaxy Package » webapps Package » galaxy Package » previous next modules index

Project Versions

latest

RTD Search

Go

Full-text doc search.

Table Of Contents

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

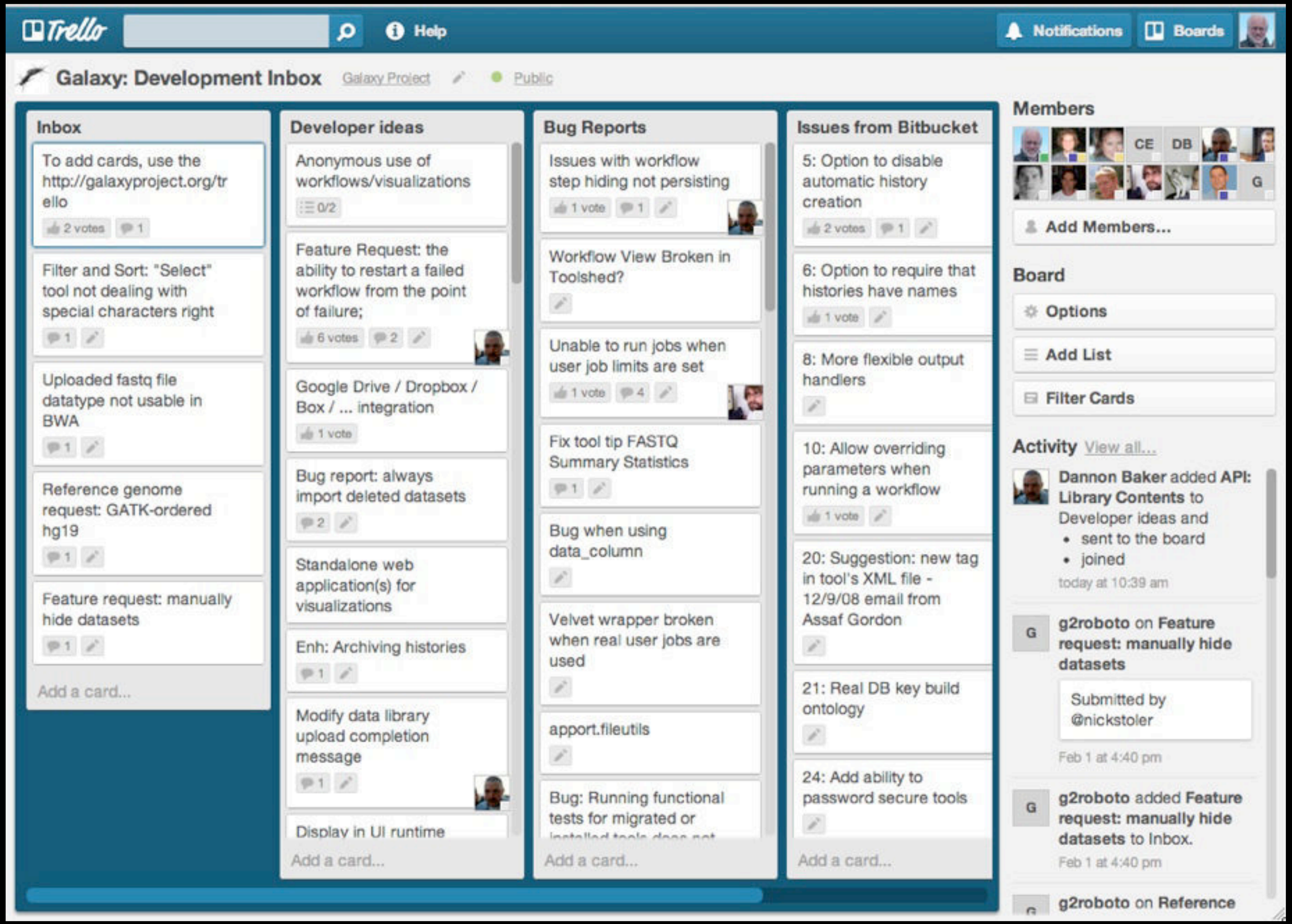
Log in as your user, navigate to the API Keys page in the User menu, and generate a new API key. Make a note of the API key, and then pull up a terminal. Now we'll use the display.py script in your galaxy/scripts/api directory for a short example:

```
% ./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
```

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


Trello
Help
Notifications
Boards
Galaxy: Development Inbox
Galaxy Project
Public
Inbox
To add cards, use the http://galaxyproject.org/trello
2 votes
1
Filter and Sort: "Select" tool not dealing with special characters right
1
Uploaded fastq file datatype not usable in BWA
1
Reference genome request: GATK-ordered hg19
1
Feature request: manually hide datasets
1
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
Google Drive / Dropbox / Box / ... integration
1 vote
Bug report: always import deleted datasets
2
Standalone web application(s) for visualizations
Enh: Archiving histories
1
Modify data library upload completion message
1
Display in UI runtime
Add a card...
Bug Reports
Issues with workflow step hiding not persisting
1 vote
1
Workflow View Broken in Toolshed?
Unable to run jobs when user job limits are set
1 vote
4
Fix tool tip FASTQ Summary Statistics
1
Bug when using data_column
Velvet wrapper broken when real user jobs are used
apport.fileutils
Bug: Running functional tests for migrated or
Add a card...
Issues from Bitbucket
5: Option to disable automatic history creation
2 votes
1
6: Option to require that histories have names
1 vote
8: More flexible output handlers
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
21: Real DB key build ontology
24: Add ability to password secure tools
Add a card...
Members
CE
DB
G
Add Members...
Board
Options
Add List
Filter Cards
Activity
View all...
Dannon Baker added API: Library Contents to Developer ideas and
sent to the board
joined
today at 10:39 am
g2roboto on Feature request: manually hide datasets
Submitted by @nickstoler
Feb 1 at 4:40 pm
g2roboto added Feature request: manually hide datasets to Inbox.
Feb 1 at 4:40 pm
g2roboto on Reference

Software

Community

Software

Community

2012 and 2013 Meetings

Galaxy Community Conference 2012

July 25-27 UIC Forum University of Illinois at Chicago

http://galaxyproject.org/GCC2012

UIC

Training Day GCC2012

Galaxy Community Conference

30 June - 2 July 2013

OSLO

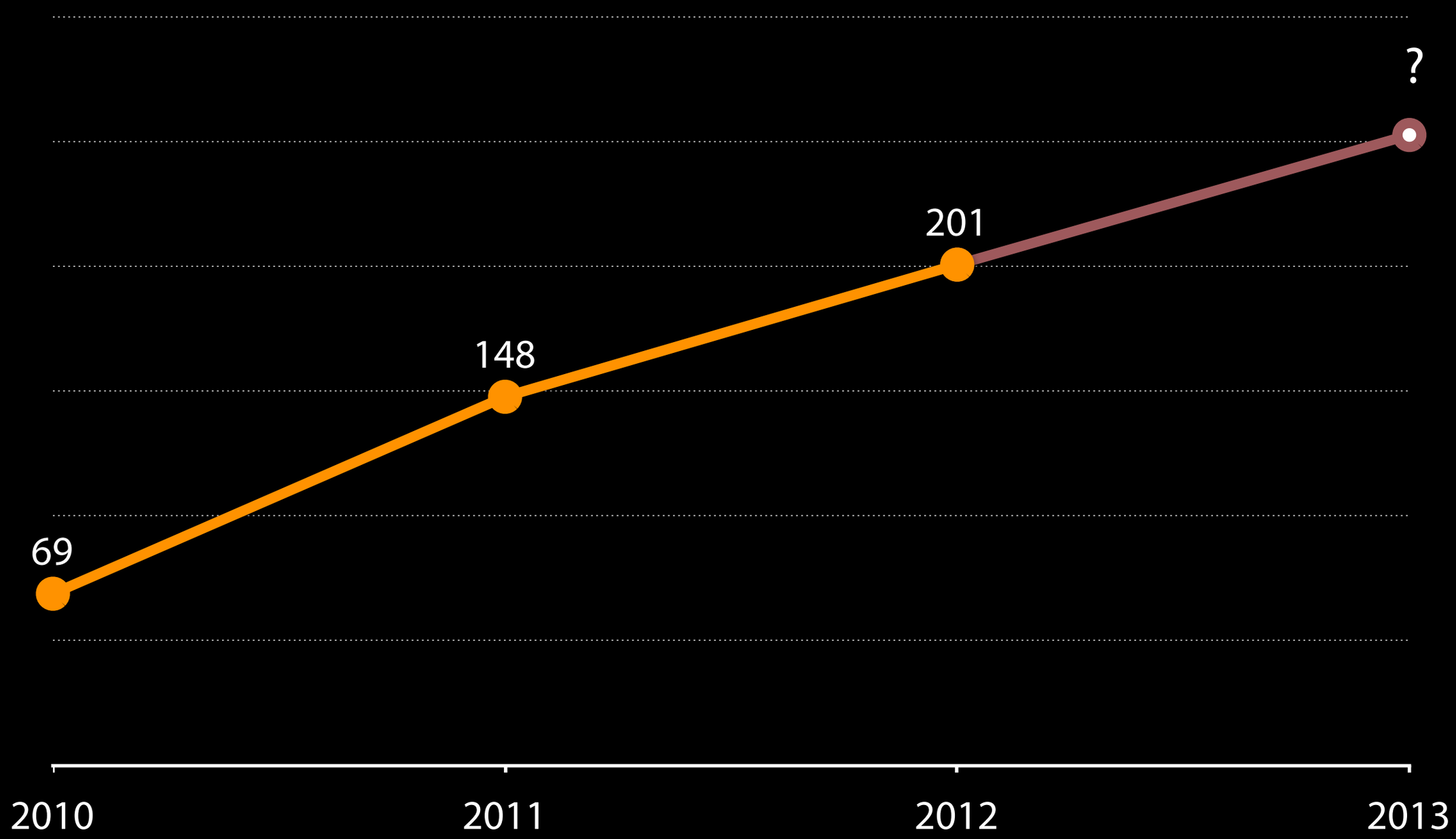
UiO : University of Oslo

GCC2013 Training Day

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

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:

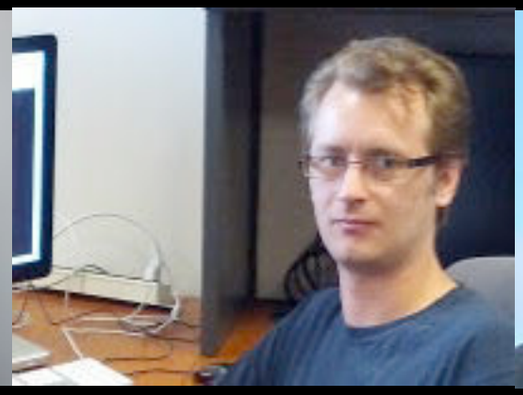
Scott Cain
Amelia Ireland

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!