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?

```python
if percentVeterans < 66:
    demoSuccess = attemptThreeMinuteDemo()

if percentVeterans >= 66 or not demoSuccess:
    handWaveOverScreenshot()
```

http://usegalaxy.org
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)

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.

**Screeencasts!**

See Galaxy Interval Operation Screeencasts (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 first dataset that overlap with the second dataset by at least the **overlap** value.

<table>
<thead>
<tr>
<th>Chrom</th>
<th>Start</th>
<th>End</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr22</td>
<td>17586742</td>
<td>17586844</td>
<td>NM_014339</td>
</tr>
<tr>
<td>chr22</td>
<td>17600280</td>
<td>17602017</td>
<td>NM_031895</td>
</tr>
<tr>
<td>chr22</td>
<td>17600280</td>
<td>17602017</td>
<td>NM_031895</td>
</tr>
<tr>
<td>chr22</td>
<td>17600280</td>
<td>17602017</td>
<td>NM_031895</td>
</tr>
<tr>
<td>chr22</td>
<td>17600280</td>
<td>17602017</td>
<td>NM_031895</td>
</tr>
</tbody>
</table>
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
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
CloudMan Platforms

GVL:
- OpenStack
- NecTAR

OpenNebula:
- NBIC server
- Andromeda
Launch a cloud instance from another running Galaxy
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 plot of 'Select first on data 1'

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

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
- Visualization framework
- ObjectStore storage API
- Galaxy API
- ...
- Make everything pluggable; start using those interfaces internally.
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 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: Unknown history
   id: 8c49be448cfe29bc
#2: /api/histories/33b43b4e7893c91f
   name: output test
   id: 33b43b4e7893c91f
```
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 Board: Galaxy Development Inbox

**Inbox**
- To add cards, use the http://galaxyproject.org/tr ello
- Filter and Sort: "Select" tool not dealing with special characters right
- Uploaded fastq file datatype not usable in BWA
- Reference genome request: GATK-ordered hg19
- Feature request: manually hide datasets

**Developer Ideas**
- Anonymous use of workflows/visualizations
- Feature Request: the ability to restart a failed workflow from the point of failure;
- Used fastq file with Google Drive / Dropbox
- Bug report: always import deleted datasets
- Standalone web application(s) for visualizations
- Enh: Archiving histories
- Modify data library upload completion message
- Evaluation of data_column
- Bug: Running functional tests for migrated or installed tools does not pass

**Bug Reports**
- Issues with workflow step hiding not persisting
- Workflow View Broken in Toolshed?
- Unable to run jobs when user job limits are set
- Fix tool tip FASTQ Summary Statistics
- Bug when using data_column
- Velvet wrapper broken when real user jobs are used
- Apport: fileutils
- Bug: Running functional tests for migrated or installed tools does not pass

**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 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

**Activity**
- View all...
- Dannon Baker added API: Library Contents to Developer ideas
- g2roboto on Feature request: manually hide datasets
- g2roboto added Feature request: manually hide datasets to Inbox.
2012 and 2013 Meetings

GCC2013 Registration & abstract submission open
http://galaxyproject.org/GCC2013
GCC attendance over time

- 2010: 69
- 2011: 148
- 2012: 201
- 2013: ?
New Communities

**GalaxyAdmins**
Administrators of large Galaxy Instances  
Started by UIowa 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**

<table>
<thead>
<tr>
<th>Month</th>
<th>January</th>
<th>February</th>
<th>March</th>
<th>April</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2 Events</td>
<td>0 Events</td>
<td>1 Event</td>
<td>7 Events</td>
</tr>
<tr>
<td></td>
<td>245 People</td>
<td>People</td>
<td>50 People</td>
<td>225 People</td>
</tr>
<tr>
<td></td>
<td>717 Participant hrs</td>
<td>Participant hrs</td>
<td>200 Participant hrs</td>
<td>1125 Participant hrs</td>
</tr>
<tr>
<td></td>
<td>Czech Rep, CA</td>
<td></td>
<td>France</td>
<td>DC, MD, IA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Month</th>
<th>May</th>
<th>June</th>
<th>July</th>
<th>August</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6 Events</td>
<td>3 Events</td>
<td>2 Events</td>
<td>1 Events</td>
</tr>
<tr>
<td></td>
<td>291 People</td>
<td>274 People</td>
<td>230 People</td>
<td>20 People</td>
</tr>
<tr>
<td></td>
<td>882 Participant hrs</td>
<td>822 Participant hrs</td>
<td>1330 Participant hrs</td>
<td>80 Participant hrs</td>
</tr>
<tr>
<td></td>
<td>France, NC</td>
<td>France</td>
<td>France, IL (GCC)</td>
<td>NC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Month</th>
<th>September</th>
<th>October</th>
<th>November</th>
<th>December</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2 Events</td>
<td>4 Events</td>
<td>3 Events</td>
<td>1 Event</td>
</tr>
<tr>
<td></td>
<td>45 People</td>
<td>102 People</td>
<td>440 People</td>
<td>50 People</td>
</tr>
<tr>
<td></td>
<td>585 Participant hrs</td>
<td>449 Participant hrs</td>
<td>720 Participant hrs</td>
<td>750 Participant hrs</td>
</tr>
<tr>
<td></td>
<td>South Africa</td>
<td>IL, IN</td>
<td>CA</td>
<td>PR</td>
</tr>
</tbody>
</table>
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!