The State of Galaxy
Galaxy timeline

2003: GALA

2005: Galaxy I

mid 2005: Galaxy is rewritten from the ground up. Pluggability is the goal.


2008: Workflows

2009: Data Libraries, Pages

2010: NGS Tools, Sharing, Tagging, Pages

2011: Visualization and visual analytics

2012: ToolShed
Publications from others on Galaxy: ~1000 @ CiteULike
36,228
0.8 Pb
Most popular NGS tools (jobs in 2012)

- bowtie (50,693)
- cufflinks (31,085)
- tophat (31,085)
- fastQC (26,217)
- MACS (26,553)
- BWA (17,465)
User Jobs per month on usegalaxy.org

January 2007 to January 2013
Plateau is due to resource constraints not reduced demand, runtime of individual jobs getting longer.
More than 30 known public Galaxy servers

Ballaxy for structure based computational biology,
Cistrome for regulatory sequence analysis,
Genomic Hyperbrowser: statistical integration of genomic data,
GigaGalaxy: integrating workflows published in GigaScience,
Pathogen Portal: comparative analysis of host response to pathogens,
...

Dozens of large scale private Galaxy instances
<table>
<thead>
<tr>
<th>Project Title</th>
<th>Institute</th>
<th>Project Number</th>
<th>PI</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENE ONTOLOGY CONSORTIUM</td>
<td>NHGRI</td>
<td>2U41HG002273-12</td>
<td>BLAKE, JUDITH A</td>
<td>JAX</td>
</tr>
<tr>
<td>A DATA COORDINATING CENTER FOR ENCODE</td>
<td>NHGRI</td>
<td>1U41HG006992-01</td>
<td>CHERRY, JOE</td>
<td>STANFORD</td>
</tr>
<tr>
<td>ADVANCING THE SCIENTIFIC POTENTIAL OF TRANSCRIPTOMICS IN AQUATIC MODELS</td>
<td>OD</td>
<td>8R24OD011199-02</td>
<td>CRESKO, WILLIAM A</td>
<td>U Oregon</td>
</tr>
<tr>
<td>ENHANCING THE GMOD SUITE OF GENOME ANNOTATION AND VISUALIZATION TOOLS</td>
<td>NHGRI</td>
<td>5R01HG004483-05</td>
<td>HOLMES, IAN H</td>
<td>UC Berkeley</td>
</tr>
<tr>
<td>STATISTICAL ANALYSIS METHODS AND SOFTWARE FOR CHIP-SEQ DATA</td>
<td>NHGRI</td>
<td>2R01HG003747-05A1</td>
<td>KELES, SUNDUZ</td>
<td>U Wisconsin</td>
</tr>
<tr>
<td>STATISTICAL POWER CALCULATIONS FOR CHIP-SEQ EXPERIMENTS</td>
<td>NHGRI</td>
<td>1R21HG006716-01</td>
<td>KELES, SUNDUZ</td>
<td>U Wisconsin</td>
</tr>
<tr>
<td>A DATA ANALYSIS CENTER FOR INTEGRATION OF FLY AND WORM MODENCODE DATASETS</td>
<td>NHGRI</td>
<td>3RC2HG005639-02S1</td>
<td>KELLIS, MANOLIS</td>
<td>MIT</td>
</tr>
<tr>
<td>COMPUTATIONAL MODELS FOR MECHANISMS OF GLOBAL TRANSCRIPTION</td>
<td>NHGRI</td>
<td>5R01HG004069-07</td>
<td>LIU, XIAOLE SHIRLEY</td>
<td>Dana-Farber</td>
</tr>
<tr>
<td>ROBUST SOFTWARE TOOLS FOR VARIANT IDENTIFICATION AND FUNCTIONAL ASSESSMENT</td>
<td>NHGRI</td>
<td>3U01HG006513-02S1</td>
<td>MARTH, GABOR T</td>
<td>Boston College</td>
</tr>
<tr>
<td>GENOMESPACE: A COMMUNITY WEB ENVIRONMENT FOR INTEGRATIVE GENOMIC ANALYSIS</td>
<td>NHGRI</td>
<td>5P01HG005062-04</td>
<td>MESIROV, JILL P.</td>
<td>Broad</td>
</tr>
<tr>
<td>UNLOCKING TRANSCRIPT DIVERSITY VIA DIFFERENTIAL ANALYSES OF SPICE GRAPHS</td>
<td>NHGRI</td>
<td>1R01HG006272-01A1</td>
<td>PRINS, JAN F</td>
<td>UNC CH</td>
</tr>
<tr>
<td>NEW ALGORITHMS AND TOOLS FOR LARGE-SCALE GENOMIC ANALYSES</td>
<td>NHGRI</td>
<td>1R01HG006693-01</td>
<td>QUINLAN, AARON R</td>
<td>UVA</td>
</tr>
<tr>
<td>INTEGRATIVE TOOLS FOR PROTOZOAN PARASITE RESEARCH</td>
<td>NIGMS</td>
<td>5R01GM093132-03</td>
<td>STOECKERT, CHRISTIAN J.</td>
<td>UPenn</td>
</tr>
<tr>
<td>PROJECT III - COMPUTATIONAL ANALYSIS OF PHENOTYPES AND</td>
<td>NICHD</td>
<td>2P01HD039691-11</td>
<td>SUCGANG, RICHARD</td>
<td>Baylor</td>
</tr>
<tr>
<td>ECOLIHub2.0: A NEXT-GENERATION E. COLI MODEL ORGANISM RESOURCE (SRI Proposal</td>
<td>NIGMS</td>
<td>5U24GM088849-04</td>
<td>THOMAS, PAUL D.</td>
<td>USC</td>
</tr>
<tr>
<td>MEGa2: MANIPULATION ENVIRONMENT FOR GENETIC ANALYSES</td>
<td>NIGMS</td>
<td>5R01GM076667-07</td>
<td>WEEKS, DANIEL E</td>
<td>Pitt</td>
</tr>
<tr>
<td>INIA: BIOINFORMATICS CORE</td>
<td>NIAAA</td>
<td>5U01AA016662-07</td>
<td>WILLIAMS, ROBERT W.</td>
<td>U Tennessee</td>
</tr>
<tr>
<td>NEW BAYESIAN ALGORITHMS FOR GENOME-WIDE ASSOCIATION MAPPING</td>
<td>NHGRI</td>
<td>2R01HG004718-04A1</td>
<td>ZHANG, YU</td>
<td>Penn State</td>
</tr>
<tr>
<td>Project Title</td>
<td>Award Number</td>
<td>PI</td>
<td>Institution</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------------</td>
<td>--------------</td>
<td>---------------------------</td>
<td>-------------------</td>
<td></td>
</tr>
<tr>
<td>GALAXY-P: A NEW COMMUNITY-BASED INFORMATICS PARADIGM FOR MS-BASED PROTEIOMICS</td>
<td>1147079</td>
<td>GRIFFIN, TIMOTHY</td>
<td>UMN</td>
<td></td>
</tr>
<tr>
<td>SOFTWARE AND SERVER FOR TAXONOMIC BINNING OF METAGENOMIC SEQUENCES</td>
<td>850103</td>
<td>TOVCHIGRENKO, ANDREY</td>
<td>JC Venter Inst.</td>
<td></td>
</tr>
<tr>
<td>GENERATION AND ANALYSIS OF A REFERENCE SEQUENCE FOR THE WEST AFRICAN CULTIVATED RICE - ORYZA</td>
<td>822284</td>
<td>WING, ROD</td>
<td>U Arizona</td>
<td></td>
</tr>
<tr>
<td>COMPUTATIONAL TOOLS AND STATISTICAL ANALYSIS OF CO-VARYING RATES OF DIFFERENT MUTATION TYPES</td>
<td>965596</td>
<td>MAKOVA, KATERYNA</td>
<td>Penn State</td>
<td></td>
</tr>
</tbody>
</table>
Developer involvement
Contributors outside the core team customize Galaxy through forks...and contribute changes back through pull requests.
21 open pull requests (as of 2/23), ~150 pull requests already merged
At least 66 distinct contributors to the main code base

Number of distinct contributors per month, gradually increasing:
<table>
<thead>
<tr>
<th>Name</th>
<th>Kudos</th>
<th>12 Month Commits</th>
<th>All Time Commits</th>
<th>5 Year Trend</th>
<th>Primary Language</th>
<th>First Commit</th>
<th>Last Commit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jeremy Goecks</td>
<td>7</td>
<td>521</td>
<td>1368</td>
<td></td>
<td>Python</td>
<td>over 3 years ago</td>
<td>4 days ago</td>
</tr>
<tr>
<td>Greg Von Kuster</td>
<td>6</td>
<td>379</td>
<td>1730</td>
<td></td>
<td>Python</td>
<td>about 6 years ago</td>
<td>4 days ago</td>
</tr>
<tr>
<td>Dan Blankenberg</td>
<td>7</td>
<td>288</td>
<td>1132</td>
<td></td>
<td>Python</td>
<td>about 6 years ago</td>
<td>5 days ago</td>
</tr>
<tr>
<td>Dannon Baker</td>
<td>7</td>
<td>157</td>
<td>349</td>
<td></td>
<td>Python</td>
<td>almost 3 years ago</td>
<td>8 days ago</td>
</tr>
<tr>
<td>Nate Coraor</td>
<td>7</td>
<td>143</td>
<td>999</td>
<td></td>
<td>Python</td>
<td>about 6 years ago</td>
<td>4 days ago</td>
</tr>
<tr>
<td>InitHello</td>
<td>6</td>
<td>133</td>
<td>133</td>
<td></td>
<td>Python</td>
<td>10 months ago</td>
<td>4 days ago</td>
</tr>
<tr>
<td>Carl Eberhard</td>
<td>6</td>
<td>128</td>
<td>128</td>
<td></td>
<td>JavaScript</td>
<td>7 months ago</td>
<td>4 days ago</td>
</tr>
<tr>
<td>jtx (Manager)</td>
<td>7</td>
<td>123</td>
<td>746</td>
<td></td>
<td>Python</td>
<td>over 6 years ago</td>
<td>15 days ago</td>
</tr>
<tr>
<td>John Chilton</td>
<td>6</td>
<td>78</td>
<td>97</td>
<td></td>
<td>Python</td>
<td>almost 2 years ago</td>
<td>8 days ago</td>
</tr>
<tr>
<td>Scott McManus</td>
<td>6</td>
<td>45</td>
<td>45</td>
<td></td>
<td>Python</td>
<td>9 months ago</td>
<td>4 months ago</td>
</tr>
<tr>
<td>Ross Lazarus</td>
<td>7</td>
<td>24</td>
<td>223</td>
<td></td>
<td>Python</td>
<td>over 5 years ago</td>
<td>4 months ago</td>
</tr>
<tr>
<td>clements</td>
<td>5</td>
<td>18</td>
<td>20</td>
<td></td>
<td>Python</td>
<td>about 1 year ago</td>
<td>about 1 month ago</td>
</tr>
<tr>
<td>Richard Burhans</td>
<td>6</td>
<td>12</td>
<td>56</td>
<td></td>
<td>XML</td>
<td>over 2 years ago</td>
<td>3 months ago</td>
</tr>
<tr>
<td>Finn Bacall</td>
<td>5</td>
<td>11</td>
<td>11</td>
<td></td>
<td>Python</td>
<td>11 months ago</td>
<td>8 months ago</td>
</tr>
<tr>
<td>Kyle Ellrott</td>
<td>1</td>
<td>11</td>
<td>11</td>
<td></td>
<td>Python</td>
<td>4 months ago</td>
<td>16 days ago</td>
</tr>
<tr>
<td>Author</td>
<td>Rank</td>
<td>5 year commit count</td>
<td>3 year commit count</td>
<td>Language</td>
<td>Last commit</td>
<td>Last commit ago</td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>------</td>
<td>---------------------</td>
<td>--------------------</td>
<td>----------</td>
<td>-------------</td>
<td>-----------------</td>
<td></td>
</tr>
<tr>
<td>Kyle Ellrott</td>
<td>1</td>
<td>11</td>
<td>11</td>
<td>Python</td>
<td>4 months ago</td>
<td>16 days ago</td>
<td></td>
</tr>
<tr>
<td>Enis Afgan</td>
<td>6</td>
<td>9</td>
<td>82</td>
<td>Python</td>
<td>over 3 years ago</td>
<td>5 months ago</td>
<td></td>
</tr>
<tr>
<td>Remi Marenco</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>XML</td>
<td>8 months ago</td>
<td>6 months ago</td>
<td></td>
</tr>
<tr>
<td>Brad Chapman</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>Python</td>
<td>over 2 years ago</td>
<td>5 months ago</td>
<td></td>
</tr>
<tr>
<td>peterjc</td>
<td>5</td>
<td>6</td>
<td>88</td>
<td>XML</td>
<td>over 2 years ago</td>
<td>4 months ago</td>
<td></td>
</tr>
<tr>
<td>Richard Park</td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>Python</td>
<td>about 1 year ago</td>
<td>24 days ago</td>
<td></td>
</tr>
<tr>
<td>Joachim Jacob</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>Python</td>
<td>about 1 month ago</td>
<td>11 days ago</td>
<td></td>
</tr>
<tr>
<td>Hiral Vora</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>XML</td>
<td>10 months ago</td>
<td>10 months ago</td>
<td></td>
</tr>
<tr>
<td>mgehring</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>Python</td>
<td>4 months ago</td>
<td>4 months ago</td>
<td></td>
</tr>
<tr>
<td>nuwan_ag</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>Python</td>
<td>6 months ago</td>
<td>6 months ago</td>
<td></td>
</tr>
<tr>
<td>Anton Nekrutenko</td>
<td>7</td>
<td>2</td>
<td>178</td>
<td>XML</td>
<td>about 6 years ago</td>
<td>10 months ago</td>
<td></td>
</tr>
<tr>
<td>Brad Langhorst</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>Python</td>
<td>7 months ago</td>
<td>7 months ago</td>
<td></td>
</tr>
<tr>
<td>Guru Ananda</td>
<td>7</td>
<td>1</td>
<td>204</td>
<td>XML</td>
<td>about 6 years ago</td>
<td>9 months ago</td>
<td></td>
</tr>
<tr>
<td>Nicola Soranzo</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>XML</td>
<td>about 1 month ago</td>
<td>about 1 month ago</td>
<td></td>
</tr>
<tr>
<td>Lance Parsons</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>Python</td>
<td>over 1 year ago</td>
<td>3 months ago</td>
<td></td>
</tr>
<tr>
<td>James</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>Python</td>
<td>6 days ago</td>
<td>6 days ago</td>
<td></td>
</tr>
</tbody>
</table>
We have also recently redesigned our release process.

All development happens in “galaxy-central”, we previously released monthly to “galaxy-dist”

However, as the number of production instances has grown, their maintainers were finding it difficult to keep up with this release cycle.

We are now releasing on a cycle of approximately three months with a longer testing period before releases.

However, we are also maintaining a dist release branch for releasing bug fixes or security fixes on the current distribution.
The path to sustainability
Leverage existing compute infrastructure

Currently engaged in several efforts to leverage national cyberinfrastructure in Galaxy

10 Gig dedicated link from Galaxy server room to XSEDE network via PSC

Mirror of all Galaxy data on PSC’s SLASH2 store

Ongoing work with XSEDE Extended Support group enabling job submission to XSEDE

Integration with Globus Online from Ian Foster’s group

Goal is to allow users to link XSEDE allocations to Galaxy accounts, submit directly from Galaxy main
Using robust infrastructure
The Galaxy Toolshed
Scaling Galaxy: two distinct problems

• So much data, not enough infrastructure.
  • Solution, encourage local Galaxy instances, cloud Galaxy, support increasingly decentralized model, *improve access to exiting resources*

• So many tools and workflows, not enough manpower
  • Focus on building infrastructure to allow community to integrate and share tools, workflows, and best practices
Galaxies on public clouds

http://usegalaxy.org
http://usegalaxy.org/toolshed

Galaxies on private clouds

private Galaxy installations

http://usegalaxy.org

Galaxy Tool Shed
http://usegalaxy.org/toolshed

1 2 3

...
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
Repositories are owned by the contributor, can contain tools, workflows, etc.

Backed by version control, a complete version history is retained for everything that passes through the toolshed.

Galaxy instance admins can install tools directly from the toolshed using only a web UI.

Support for recipes for installing the underlying software that tools depend on (also versioned).

Wednesday, July 17, 13
Toolshed oversight committee

- Ira Cooke | La Trobe
- JJ Johnson | UMN
- Ed Kirton | DOE JGI
- Peter Cock | Hutton Institute
- Brad Chapman | HSPH
- Björn Grüning | Freiburg
- Ross Lazarus | Baker IDI
Improving robustness and performance of the Galaxy UI
Although the core of the Galaxy framework is robust and scalable, there are some areas we need to invest engineering effort in revisiting.

Some features currently perform poorly due to inefficient use of the database.

Many components communicate through synchronization on the database, need to refactor the framework with a proper message bus.
Supporting analysis at massive scale
One of the most common requests from users is to run tools or workflows on many samples.

For example, a user might have a few dozen (paired) samples, want to run each 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.
Limitations of the Galaxy Workflow System

It isn’t a workflow system

Launching large workflows is relatively slow

Scheduling is completely naive
(just scheduled as individual jobs)

Not robust to individual job failures, no way to
detect types of failures or rerun individual jobs or
parts of workflows
Plans

Schedule workflows in the background
Histories will be able to contain not just datasets, but pending workflows, dataset groups, other entities

Make workflow scheduling engine pluggable
Once it is a background process, can afford the time to delegate

Rewrite default workflow scheduling approach
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

Better integration of workflow planning with job level parallelism
Limitations of the Galaxy user interface

Very good at what is was designed for, running one tool, or one workflow, on a few datasets interactively

Very limited facilities for running on multiple datasets
Running workflow "Groom FASTQ"

Step 1: Input dataset

Input Dataset
2: imported: GM12878ple Dataset

Step 2: FASTQ Groomer (version 1.0.4)

Send results to a new history
Run workflow
Running workflow "Groom FASTQ"

Step 1: Input dataset

Input Dataset

1: Imported: h1-hESC..ple Dataset
2: Imported: GM12878..ple Dataset

type to filter, [enter] to select all

Step 2: FASTQ Groomer (version 1.0.4)

Send results to a new history

Run workflow
Need to support much more complex combinations of multiple inputs

Parallel execution on some datasets, cross products on others, match up paired datasets, annotations, ...

Need to provide better feedback to users running large-scale analysis on shared resources
Plans

Completely new view in Galaxy for multiple dataset execution

Given a workflow to run, use rules to select sets of datasets that it will be run on

Define tools or workflows for aggregating results

Completely new interfaces for monitoring progress of running workflows and jobs, not limited to single history

Better feedback on queue position and resource utilization

Wednesday, July 17, 13
Galaxy Team Roadmap for 2014

- ToolShed
- Cloud
- Improving data integration & management
- User Interface refactoring/redesign
- Visualization framework and API
- Robust federated infrastructure for main site
- Working more closely with journals
This will not be possible without

**Organizing Committee**
- Nils Christophersen
- Dave Clements
- Ingeborg Engh
- Mohammad Heydarian
- Geir Kjetil Sandve

**BoF Makers**
- Dave Clements
- John Chilton
- Ira Cooke
- Björn Grüning
- JJ Johnson
- Peter Cock
- Sanjay Joshi
- Ross Lazarus

**Session Moderators**
- Nils Christophersen
- Victoria Dominiguez
- Mohammad Heydarian
- Ross Lazarus
- Karen Reddy
- Jennifer Hillman-Jackson
- Geir Kjetil Sandve
- Hans-Rudolf Hotz

**Scientific Committee**
- Nils Christophersen
- Karen Reddy
- Geir Kjetil Sandve

**Training Day Instructors**
- Nate Coraor
- Geir Kjetil Sandve
- Sveinung Gundersen
- Katerina Michalickova
- Nikolay Aleksandrov Vazov
- Jennifer Jackson
- Enis Afgan
- Dannon Baker
- Carl Eberhard
- Dan Blankenberg
- Ross Lazarus
- Jeroen F.J. Laros
- Wibowo Arindrarto
- Leon Mei
- Greg Von Kuster
- Erik Garrison
- Alban Lermine
- Valentina Boeva

Wednesday, July 17, 13
The Galaxy Team

http://wiki.galaxyproject.org/GalaxyTeam

Wednesday, July 17, 13