

Introduction to Galaxy

University of California,
Berkeley

The logo for the Galaxy project, featuring a stylized icon of three horizontal bars of varying lengths (white, grey, yellow) to the left of the word "Galaxy" in a white sans-serif font.

7 November 2012

Jeremy Goecks
Emory University

Jennifer Jackson
Penn State University

galaxyproject.org

Agenda

- 9:00 **Welcome, Introduction to Galaxy**
- 9:15 Galaxy 101 Tutorial: Data, Tools, Workflows
- 10:15 *Break*
- 10:30 RNA-Seq, Visualization and Visual Analytics
- 11:30 Galaxy Usage, Support, and Resources
- 11:45 Open discussion
- 12:00 *Done*



Enis Afgan



Guru Ananda



Dannon Baker



Dan Blankenberg



Dave Bouvier



Dave Clements



Nate Coraor



Carl Eberhard



Jeremy Goecks



Nuwan Goonasekera



Jen Jackson



Greg von Kuster



Ross Lazarus



Rémi Marengo



Scott McManus



**Anton
Nekrutenko**



**James
Taylor**

The Galaxy Team

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

Goals for this workshop

Introduction

- What's Galaxy? And some Core **Terminology**

Hands-on experience

- **Perform bioinformatics analysis with Galaxy**
- **Not only visualize** your results, but learn how **visual analytics can empower your research**

Moving forward

- Learn **how to stay connected to the Galaxy community for support** once you begin your **own projects**
- Where to find more **tutorials** and advanced **'live supplementals'**.
- How to **describe, share,** and **publish** results.
- **Galaxy Options:** Local, Cloud, Other Galaxies ...

Today is just the start

This workshop is about “Using Galaxy” and it will not cover in depth details about how the tools are implemented or new algorithm designs or which assembler or mapper or X ...

... is best for you.

BUT, we *really* like to help. **Tool forms are packed with useful information.** And you can reach out during the break. If we don't have time today, know that **all material is available online** and you can **review our support wiki** anytime. If you get stuck or need more info, **send an email to one of our mailing lists.**

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

So, ...What is Galaxy?

- A **data analysis and integration** tool
- 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
- These options result in several **ways to use Galaxy**

<http://galaxyproject.org>

Galaxy 101 Terminology

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

Agenda

- 9:00 Welcome, Introduction to Galaxy
- 9:15 **Galaxy 101 Tutorial: Data, Tools, Workflows**
- 10:15 *Break*
- 10:30 RNA-Seq, Visualization and Visual Analytics
- 11:30 Galaxy Usage, Support, and Resources
- 11:45 Open discussion
- 12:00 *Done*

Agenda

- 9:00 Welcome, Introduction to Galaxy
- 9:15 Galaxy 101 Tutorial: Data, Tools, Workflows
- 10:15 *Break*
- 10:30 **RNA-Seq, Visualization and Visual Analytics**
- 11:30 Galaxy Usage, Support, and Resources
- 11:45 Open discussion
- 12:00 *Done*

Agenda

- 9:00 Welcome, Introduction to Galaxy
- 9:15 Galaxy 101 Tutorial: Data, Tools, Workflows
- 10:15 *Break*
- 10:30 RNA-Seq, Visualization and Visual Analytics
- 11:30 **Galaxy Usage, Support, and Resources**
- 11:45 Open discussion
- 12:00 *Done*

Galaxy URLs to Remember

<http://galaxyproject.org>

<http://usegalaxy.org>

<http://getgalaxy.org>

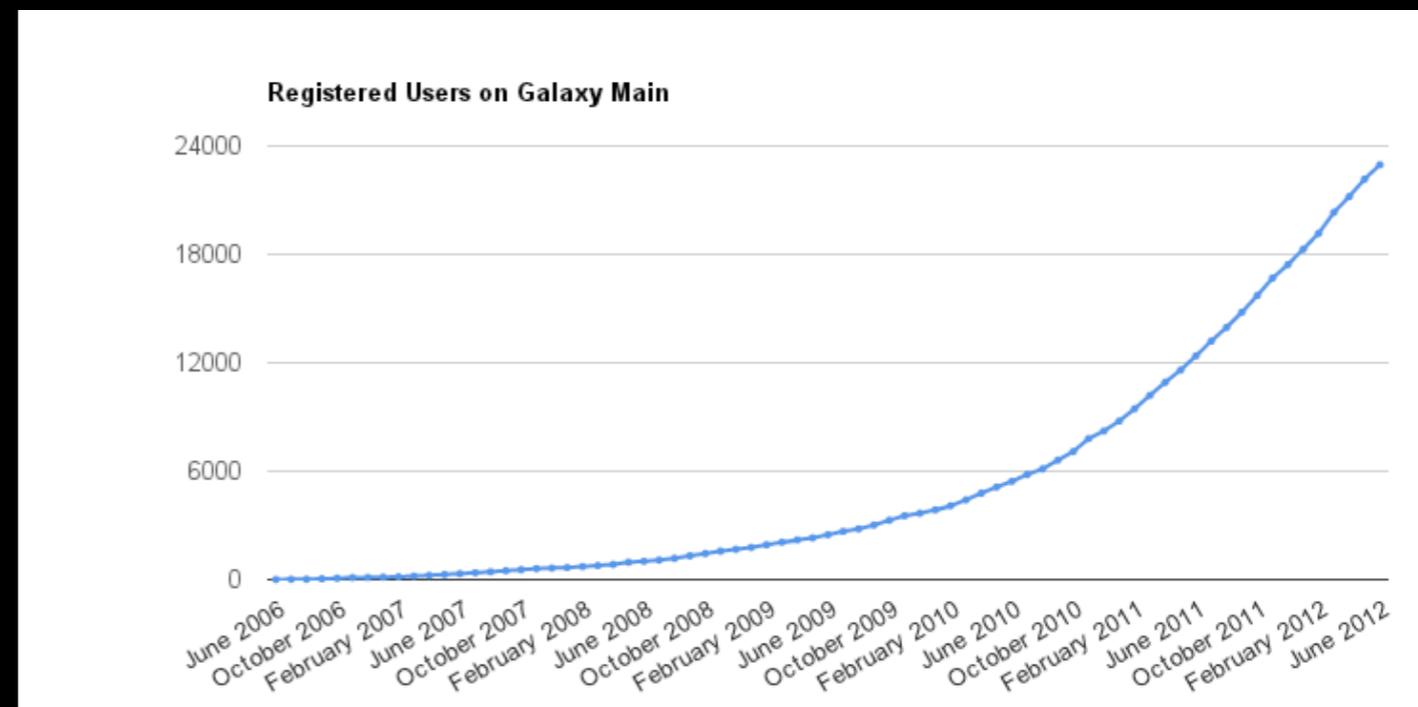
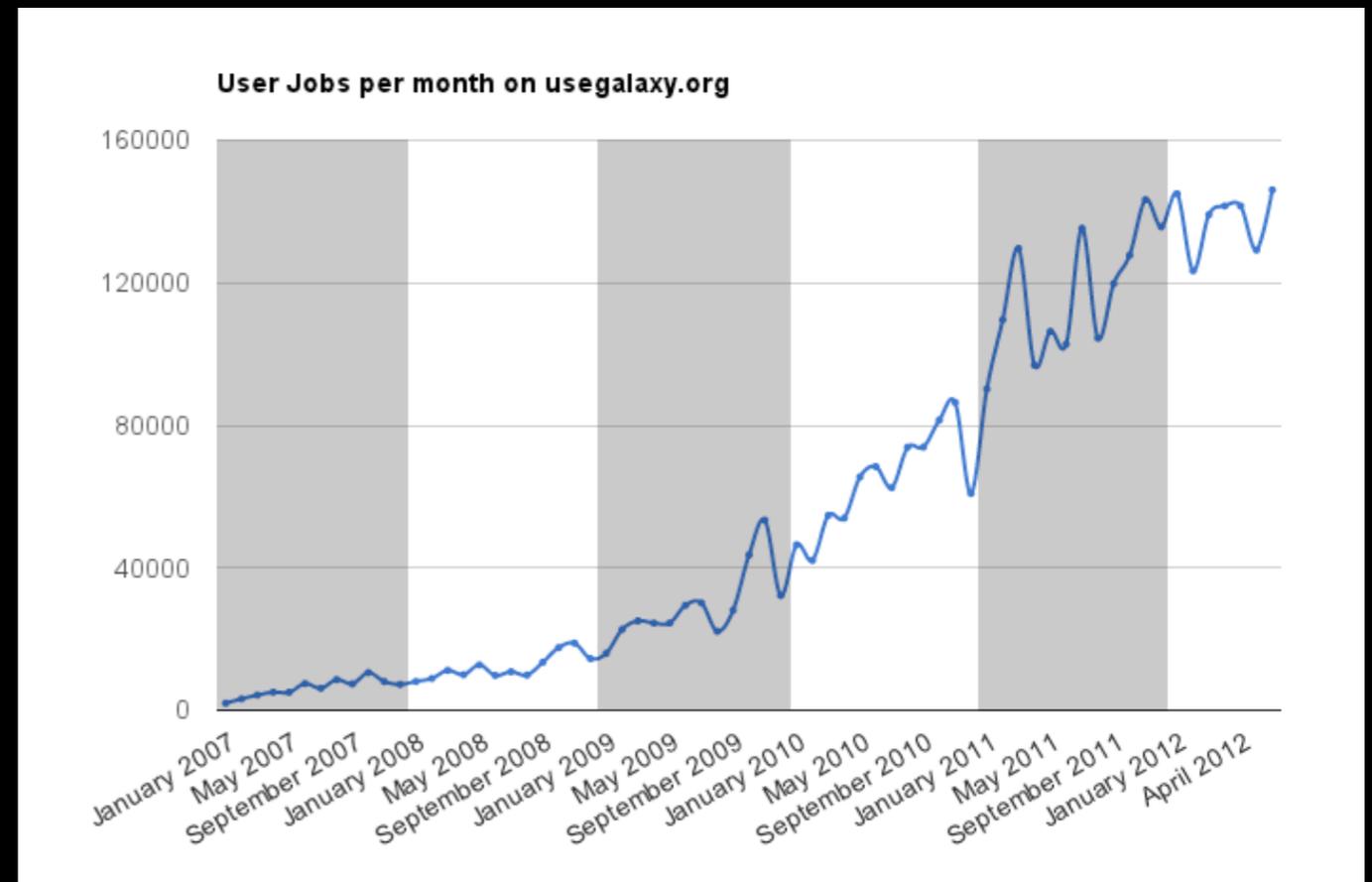
So, ...What is Galaxy? Redux

- A **data analysis and integration** tool
- 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
- These options result in several **ways to use Galaxy**

<http://galaxyproject.org>

<http://usegalaxy.org> (a.k.a Main)

- **Public web site**
- **Anybody can use it**
- Hundreds of tools
- **Persistent**
- + 500 users / month
- ~100 TB of user data
- ~140,000 jobs / month



<http://bit.ly/gxystats>

Sharing and Publishing Your Work

The screenshot shows the top of a web page for Genome Research. The logo for CSH PRESS GENOME RESEARCH is on the left. To the right is an advertisement for Illumina with the text 'Apply today for the Cancer GWAS Grant.' Below the logo is a navigation menu with links: HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CONTACT | HELP. A blue bar contains the text 'Institution: PENN STATE UNIV Sign In via User Name/Password' and a search box with 'Search for Keyword: Go' and 'Advanced Search'.

The main content area features the article title 'Windshield splatter analysis with the Galaxy metagenomic pipeline' by Sergei Kosakovsky Pond and Samir Wadhawan. A section titled 'OPEN ACCESS ARTICLE' contains publication details: 'This Article Published in Advance October 9, 2009, doi: 10.1101/gr.094508.109 Copyright © 2009 by Cold'. To the right is a 'Current Issue' section for October 2010, 20 (10).

A 'Footnotes' section is highlighted with an orange oval. The footnote text is: '[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>

But, it's a big world

Main has lots of tools, storage, processor, users, ...

- But **not all tools** - there are thousands and adding new tools is not taken lightly
- But **not infinite storage and processors** - Main now has job limits and storage quotas

A centralized solution cannot scale to meet data analysis demands of the whole world

Scaling Galaxy & Options

- **Encourage local Galaxy instances and Galaxy on the cloud**
- Support **increasingly decentralized model** and *improve access to existing resources*
- Focus on building **infrastructure to enable the community to integrate and share** tools, workflows, and best practices
- Support Community & Alternate Public Galaxy

Local Galaxy Instances

<http://getgalaxy.org>

Galaxy is designed for local installation and customization

- Easily integrate new tools
- Easy to deploy and manage on nearly any (Unix) system

Got your own cluster?

- Move tool execution to other systems
- Galaxy works with any DRMAA compliant cluster job scheduler (which is most of them).
- Galaxy is just another client to your scheduler.



Galaxy Tool Shed

- Allow users to share “suites” containing tools, datatypes, workflows, sample data, and automated installation scripts for tool dependencies
- Integration with Galaxy instances to automate tool installation and updates

toolshed.g2.bx.psu.edu

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.
- **We are using this today**



<http://aws.amazon.com/education>

Instant CloudMan

The screenshot shows the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User'. A 'Using 0%' indicator is in the top right. On the left, a 'Tools' sidebar is visible with a search bar and a list of tools under 'Get Data'. The main content area displays 'Managing Data' with the text 'Store, Manage, and Share data with Libraries' and 'An in-depth tutorial'. A 'Live Quickies' section is partially visible at the bottom. A tooltip for 'New Cloud Cluster' is shown over the 'Cloud' menu item. On the right, a data history panel shows '0 bytes' and a message: 'Your history is empty. Click 'Get Data' on the left pane to start'.

Launch a CloudMan instance directly from Main, and transfer your current history.

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

- Cluster Name:
- Password:
- Key ID:
- Secret Key:
- Instance Share String (optional):
- Instance Type:

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.

Public Galaxy Servers

Interested in:

ChIP-chip and ChIP-seq?

✓ Cistrome

Statistical Analysis?

✓ Genomic Hyperbrowser

Sequence and tiling arrays?

✓ Oqtans

Text Mining?

✓ DBCLS Galaxy

Reasoning with ontologies?

✓ GO Galaxy

Internally symmetric protein structures?

✓ SymD

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

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

CiteULike group, Mendeley mirror

Annual **Galaxy Community** Meeting

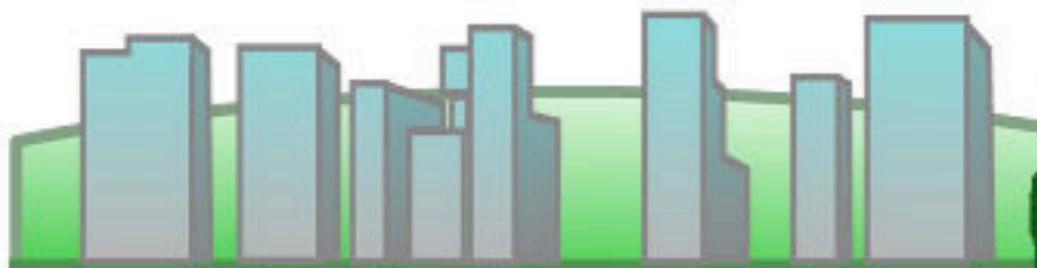
<http://galaxyproject.org/wiki>

Galaxy

Community Conference

30 June
- 2 July

2013



OSLO



UiO : University of Oslo

<http://galaxyproject.org/GCC2013>

Agenda

- 9:00 Welcome, Introduction to Galaxy
- 9:15 Galaxy 101 Tutorial: Data, Tools, Workflows
- 10:15 *Break*
- 10:30 RNA-Seq, Visualization and Visual Analytics
- 11:30 Galaxy Usage, Support, and Resources
- 11:45 **Open discussion**
- 12:00 *Done*

Agenda

- 9:00 Welcome, Introduction to Galaxy
- 9:15 Galaxy 101 Tutorial: Data, Tools, Workflows
- 10:15 *Break*
- 10:30 RNA-Seq, Visualization and Visual Analytics
- 11:30 Galaxy Usage, Support, and Resources
- 11:45 Open discussion
- 12:00 ***Done ... Thanks for using Galaxy !!!***

Acknowledgements

Madhavan Ganesh

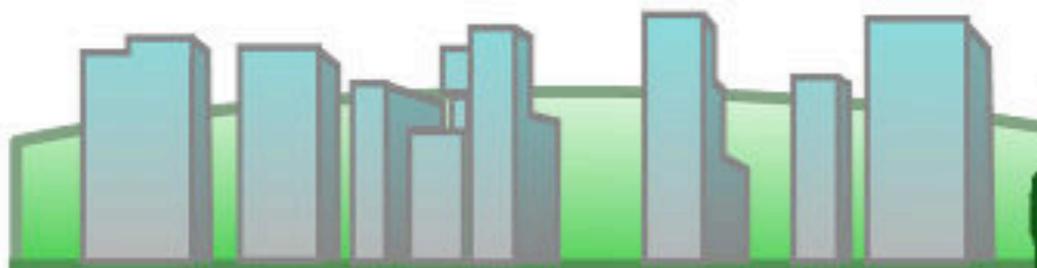
QB3-CGRL and EBI, UC Berkeley

Galaxy

Community Conference

30 June
- 2 July

2013



OSLO



UiO : University of Oslo

<http://galaxyproject.org/GCC2013>