Galaxy Project Update

5th Edinburgh Bioinformatics Meeting
University of Edinburgh
12 May 2014

Dave Clements
Johns Hopkins University

http://galaxyproject.org

edinburgh genomics
design execution analysis

igmm
INSTITUTE OF GENETICS & MOLECULAR MEDICINE

ROSLIN

Wellcome Trust Centre for Cell Biology WTCCB

Edinburgh Bioinformatics

THE UNIVERSITY of EDINBURGH

Agenda

- Project Introduction (brief)
- Project Update

What is Galaxy?

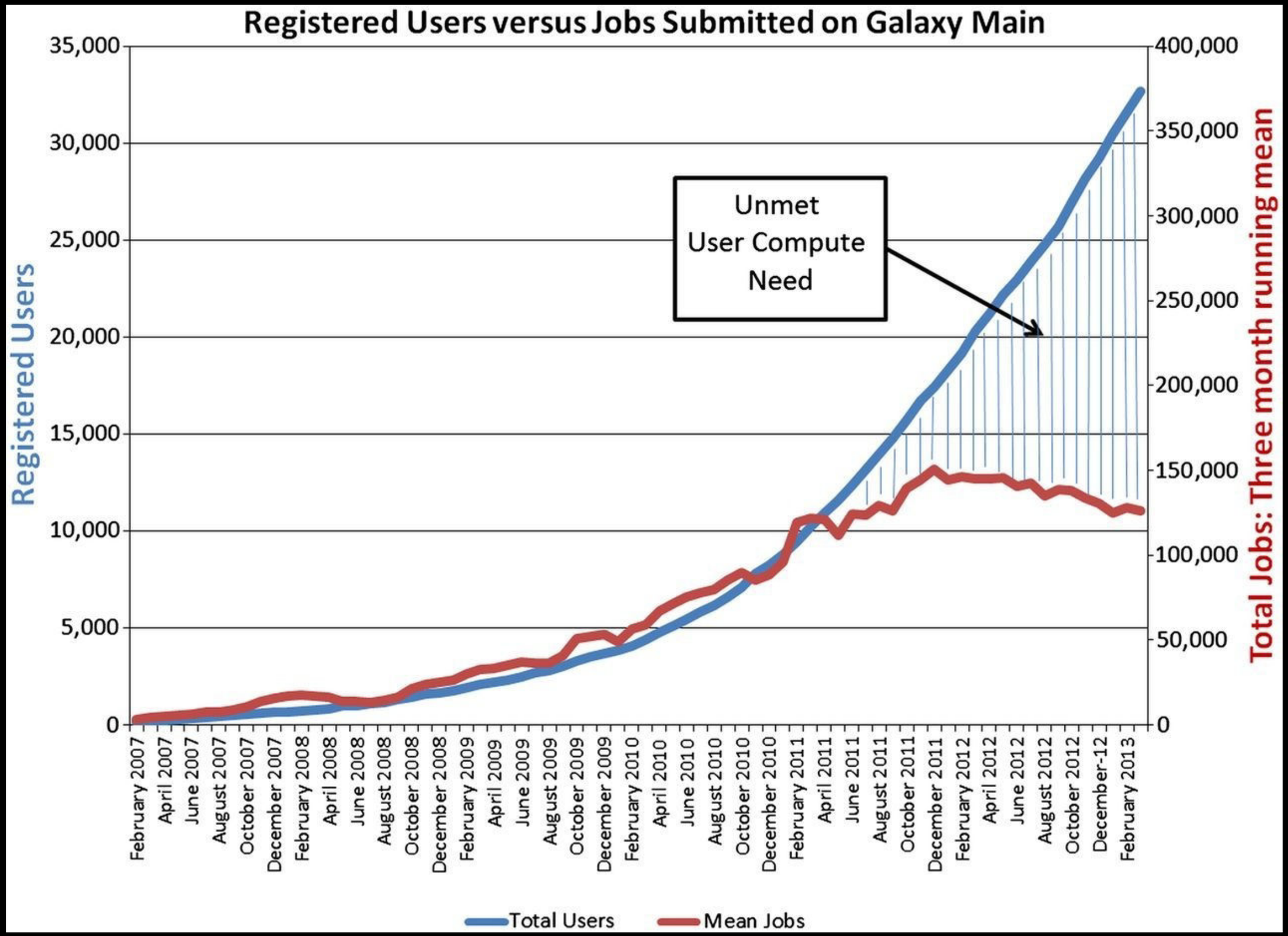
- A web based data integration and analysis framework.
- **Open source software**

http://galaxyproject.org

Galaxy is available as: usegalxy.org

The Galaxy Project's free for everyone web server integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage.

However, ***a centralized solution cannot support the different analysis needs of the entire world.***



Leveraging the national cyberinfrastructure for biomedical research

LeDuc, *et al. J Am Med Inform Assoc doi:10.1136/amiajnl-2013-002059*

Galaxy is available as

- As a free (for everyone) web service

  http://usegalaxy.org

- As open source software

  http://getgalaxy.org

It is installed at locations around the world

Galaxy is available on the Cloud

CloudMan

Start with a **fully configured & 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://aws.amazon.com/education
http://globus.org/
http://wiki.galaxyproject.org/Cloud

amazon web services™

openstack™
CLOUD SOFTWARE

OpenNebula.org
The Open Source Toolkit for Cloud Computing

globus genomics

Galaxy is available: With Commercial Support

A ready-to-use appliance
(BioTeam)

Cloud-based solutions
(ABgenomica, AIS, Appistry, GenomeCloud)

Consulting & Customization
(Arctix, BioTeam, Deena Bioinformatics)

SLIPSTREAM APPLIANCE GALAXY EDITION
Galaxy made easy.

abgenomica

GenomeCloud

Appistry

AIS™ ENTERPRISE-CLASS DATA CENTER SERVICES

Agenda

- Project Introduction
- Project Update

New Development

- Tool Shed Work
  - Lots of enhancements to handle dependencies.
  - Continue moving tools out of source distribution and into Tool Shed

New Development

- User Management
  - Disk quotas added a few years ago
  - Limited execution concurrency before that
    - Led to widespread abuse w/ multiple accounts
- Put effort into ending that on Main
  - Email verification and duplicate detection
  - Added supporting code into distribution

Release Cycle: Galaxy

Experimented with 2-3 week release cycle

Less than popular

Aiming for every 2 months or so

2013/04/01 Release

2013/06/03 Release

2013/08/12 Release

2013/11/04 Release

2014/02/10 Release

2014/04/14 Release

Releases

Project now makes extensive use of Trello cards for tracking work. This is reflected in release notes:

Core

1. Explicitly set `TEMP` dir in `Local Runner`, when a temp dir value is not already set. https://trello.com/c/HbFeoWRI
2. Tool element return_code (under stdio) now functions from_work_dir or when setting metadata externally. https://trello.com/c/JfB2w1Br
3. Using Auto-detect and a cluster job runner now sets metadata only once. https://trello.com/c/Kc3NDGyN
4. Upgrades to `HierarchicalObjectStore`, more planned. https://trello.com/c/k4tovlFd
5. New `Plugin Framework` `lib/galaxy/web/base/pluginframework.py`. https://trello.com/c/IrfWbtw3
6. Plugins define hook functions called by a Galaxy app when certain events/situations happen. https://trello.com/c/c2AzV3Xf

Releases

Doing a much better job of incorporating pull requests from community into dist:

Pull Requests Merged

1. Björn Grüning contributed a method to implement the ability to change the tool-panel as user preference (`Dynamic Toolbox Filtering`). #179. This was a frequently requested feature by the community and full documentation on this can be found here UserDefinedToolboxFilters. https://trello.com/c/XI7CZFMd
2. Björn Grüning also contributed several extensions allowing developers to utilize new actions simplifying various tool shed dependency definition idioms:
   - `make_install` action. #217
   - `autoconf` action. #218
   - `setup_r_environment` action. #219 Further extensions enhancing this last tag and a corresponding setup_ruby_environment tag from Björn will be forthcoming in the next release.
3. Additionally, Björn Grüning contributed other tool shed and tool related enhancements enhancements: #205, #216, and #239
4. Andrew Warren contributed an API method allowing coping datasets between histories as well as support for more secure e-mail settings. #199 and #198.
5. Nicola Soranzo contributed small fixes for various tools as well as enhancements for customizing and localizing data and time display in various parts of Galaxy. #222 and #211.
6. Kyle Ellrott contributed many enhancements for the API and the Galaxy search engine. #187, #241, and #234.
7. Lance Peterson contributed two enhancements to management scripts. #196 and #158. https://trello.com/c/qzjBuljp
8. Google Summer of Code Intern Saket Choudhary contributed enhancements for VCF 4.1 compatibility. #184.
9. Matthew Shirley contributed grammar fixes to the tool shed interface. #210.
10. Stephen Mcmahon contributed fixes to the PBS job runner's staging functionality. #194
11. Rémy Dernat contributed enhancements to the administrative interface allowing for management of user API keys. #134
12. Adam Brenner contributed an enhancement making it easier to deploy the histogram2 tool. #215.
13. A. Rretaud contributed extensions enabling data source tool developers to utilize the tool runners login e-mail address when implementing such tools. #206
14. John Chilton fixed job splitting to rewrite references in config files in addition to command-line. #169. https://trello.com/c/FMPydE8L
15. John Chilton and Simon Guest implemented configurable plugins for tool dependency resolution. #228. https://trello.com/c/cP3tGSJv
16. John Chilton implement `GALAXY_SLOTS` allowing tools to uniformly obtain allocated thread count. #236. https://trello.com/c/cfOlSfdP
17. Kyle Ellrott contributed enhancements that allow API tool's `POST` to define history for tool state. #193. https://trello.com/c/hpFanyx0

Release Cycle: CloudMan

After a lull, now doing semi-annually

2013/07 Release

2014/01 Release

Where are we going?

Dataset Collections

Support dataset collections as 1st class objects.

Run tools once on each dataset in the collection.

Run tools on the collection as a whole.

Tools become much more dynamic, flexible and responsive to input.

Support map/reduce paradigm.

Makes it possible to build workflows that can reason about paired datasets, technical replicates, multiple biological samples, ...

Big Data: Plans

Rewrite default workflow engine

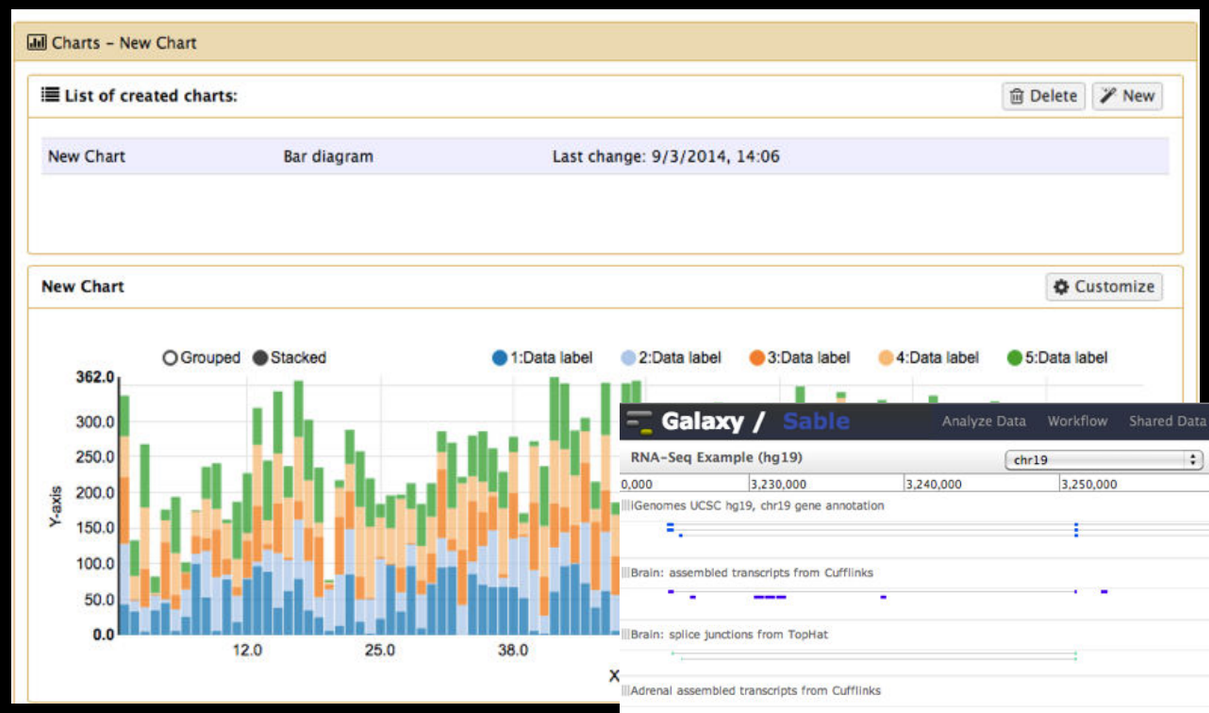
Histories will be able to contain pending workflows, dataset groups, other entities - not just datasets

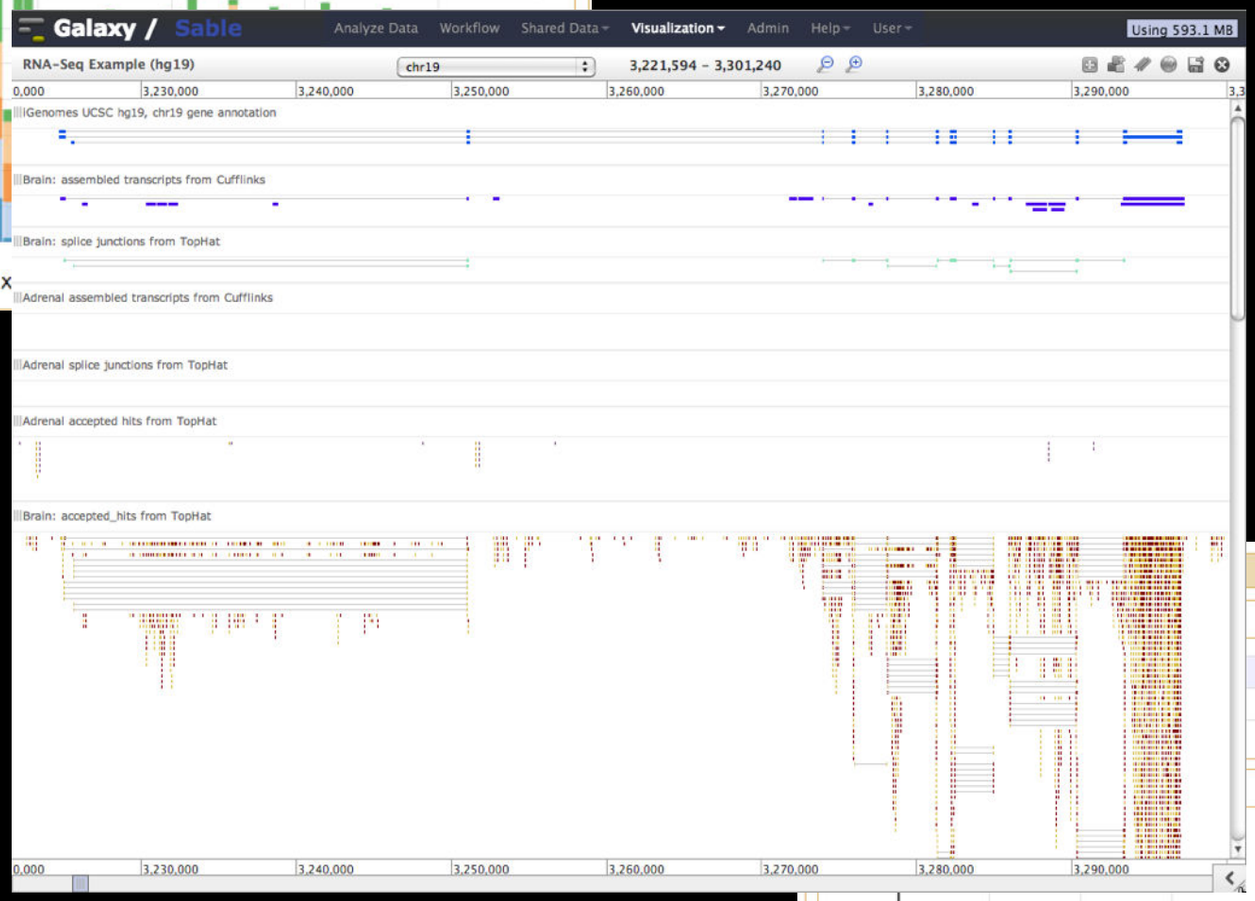
Rather than scheduling all at once, monitor workflow progress, allow pausing in response to failure or user intervention, decision nodes, streaming data and intermediate datasets, ...

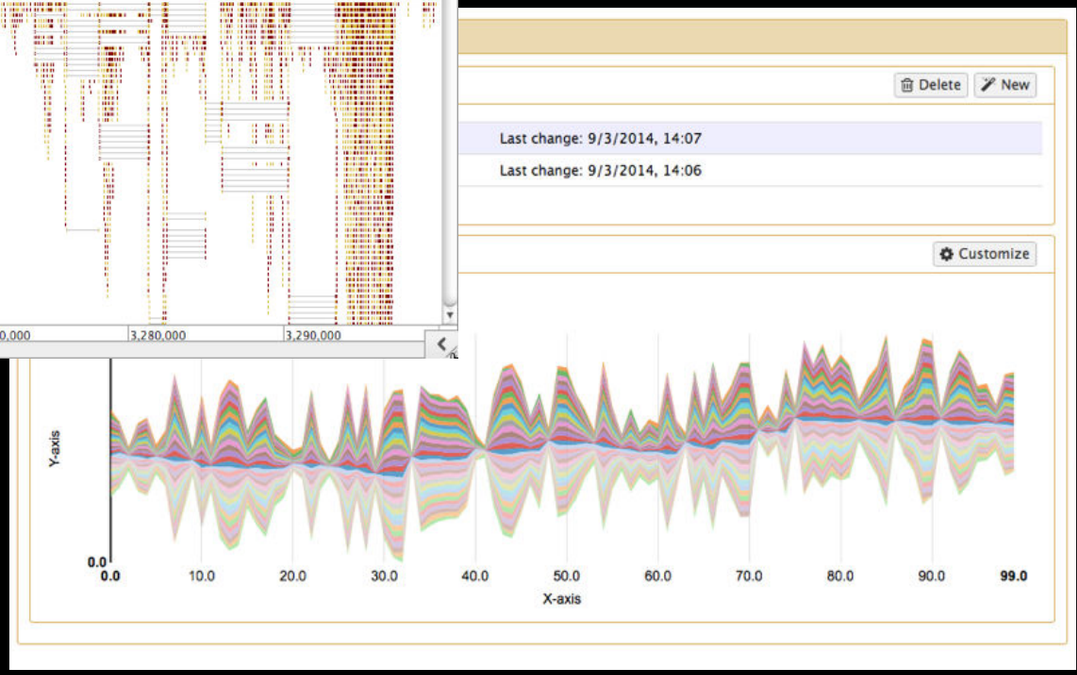
Make workflow scheduling engine pluggable

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

Visual Analytics





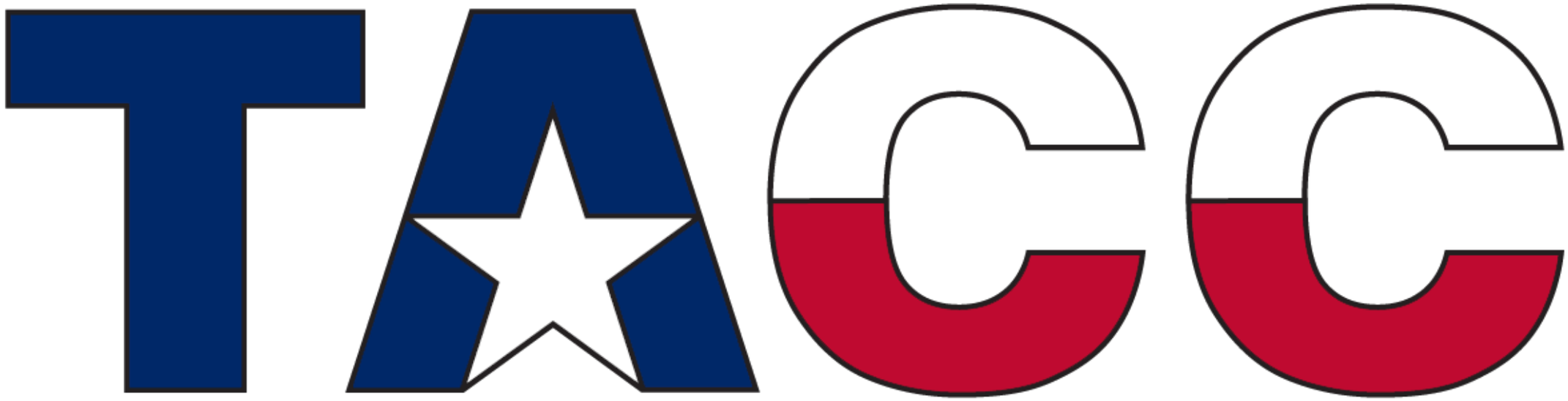


Pluggability / Extensibility / APIs

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

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 federation
- Intergalactic Utilities Commission



usegalaxy.org moved to the Texas Advanced Computing Center on October 8

But it is still not enough to
meet the analysis needs of the world.

The Galaxy Project continues to emphasize

Cloud Installs
Local Installs
Public Servers

Community

Community Hubs

Galaxy Deployment Catalog

Welcome to the *Galaxy Deployment Catalog*. This catalog describes the details of how Galaxy is installed at different institutions. Details include infrastructure information as well as user community and domain information for each deployment.

You are strongly encouraged to add your local Galaxy deployment to this catalog so that the Galaxy Community can benefit from your experience.

Contents

Deployments: User and Domain Information

Deployment	Domain	Owners	User Base	Audience
CSIRO Galaxy Service	NGS analysis: RNAseq, Genomics, metagenomics, custom tools.	CSIRO	209 registered users and 30 active users	
Galaxy server at the FMI	NGS analysis, MA analysis, custom tools.	Friedrich Miescher Institute for Biomedical Research		wet lab scientists
GalaxyAtUIowa	Local instance at the U of Iowa	IIHG		Human geneticists, biologists
Idaho State University MRCF	Bioinformatics, biology, next generation sequencing. Custom tools.	Idaho State University Molecular Research Core Facility (MRCF)		Idaho State University and Molecular Research Core Facility (MRCF) customers.
Sigenae Bioinfo-Genotoul	Genomics, esp. read alignment, SNP calling and annotation, RNA-Seq, and sRNAseq	Sigenae Team, GenoToul Bioinfo		bioinfo gentoul and INRA researchers and collaborators
URGI	Genomics	INRA	~ 30	researchers at the institution, lab members
UAB Galaxy	NGS analysis: Genomics (reference and de novo), RNAseq, metagenomics, custom tools.	Collaboration between Center for Clinical and Translational Science and Research Computing	currently ~200	Informatics core, sequencing core, researchers and students at the institution
ZBIT	Bioinformatics: SBML tools,	Center for Bioinformatics Tuebingen		

GALAXY COMMUNITY CONFERENCE
BALTIMORE, MD | JUNE 30 - JULY 2, 2014
Training Day voting closes Friday

Use Galaxy
Servers • Learn
Main • Share • Search

Communicate
Support • News
Events • Twitter
Mailing Lists (search)

Deploy Galaxy
Get Galaxy • Cloud
Admin • Tool Config
Tool Shed • Search

SLIPSTREAM APPLIANCE Galaxy made easy.

Contribute
Tool Shed • Share
Issues & Requests
Teach • Support

Galaxy Project
Home • About
Community

Community Hubs

Galaxy Wiki

Login | Search: Titles Tex

Community/Deployment/GalaxyAtUIowa

Locked History Action

THE UNIVERSITY OF IOWA

GalaxyAtUIowa

This is a local installation of Galaxy at the University of Iowa.

Deployment
GalaxyAtUIowa

Domain
Local instance at the U of Iowa

Owners
IIHG

Audience
Human geneticists, biologists

User Base

Server Topology
SGE cluster

Compute

Memory

Storage
NFS

Disk Space

User Management
University or IIHG affiliation required

Domain

Human genetics, biology. Custom tools and reports.

Community

The University of Iowa community and members of the Iowa Institute of Human Genetics.

Compute Infrastructure

This instances of Galaxy runs on a local HPC cluster.

Compute

3600 cores, SGE DRM.

Storage

GALAX
COMMUNIT
CONFERENC
BALTIMORE, MD | JUNE 30 - JULY 2, 20

Training Day voting closes Friday

Use Galaxy

Servers • Learn
Main • Share • Search

Communicate

Support • News
Events • Twitter
Mailing Lists (search)

Deploy Galaxy

Get Galaxy • Cloud
Admin • Tool Config
Tool Shed • Search

SLIPSTREAM APPLIANCE
Galaxy made easy.

Contribute

Tool Shed • Share
Issues & Requests

Community Hubs

Galaxy Community Log Board

The *Galaxy Community Log Board* is a place to share how you addressed a particular task in your Galaxy deployment. Log entries describe specific solutions to particular tasks, such as the details of what steps were taken to deploy Galaxy on particular platforms, or with particular software. It's an easy way to help others (and learn from others) by sharing what you've already done.

So, if you have figured out how to do something, and it took you a while, then this is the ideal forum for sharing that information with the community.

Contents

Logs

2014

Date	Topic	Resolution	Who
2014/03/20	**Basic Galaxy Puppet Module**	A puppet module for a very basic galaxy server (use for development)	Olivier Inizan, Mikael Loaec
2014/01/27	**Problem with logout when using LDAP for authentication with remoteUser enabled.**	Make changes to 4 Galaxy configuration files	Tim Booth

2013

Date	Topic	Resolution	Who
2013/12/05	**Tool Integration Short Tutorial**	Documents best tool integration practices from Institut Français de Bioinformatique Galaxy working group	Contributors
2013/11/13	**VelvetG error on CloudMan instance: cannot find 'cov_cutoff'**	Fixed XML wrapper	Dave Clements
2013/10/27	**IGV Integration**	How to set up IGV-Galaxy integration in Apache and Galaxy, including adding custom genomes.	Sarah Maman, Nabihoudine Ibou

Community: Public Galaxy Instances

http://bit.ly/gxyServers

Passed ~~50~~ 60 this year

- OSDD
- CoSSci
- Galaxy-P
- GO Galaxy
- Orione

Public Tool Sheds

- GenOuest
- DTL

Upcoming Events

In the next 60 days: Norway, France, online, Italy, Croatia, the Netherlands.

Date	Topic/Event	Venue/Location	Contact
May 6-7	*Scaling Galaxy for Big Data*	NGS Data after the Gold Rush, TGAC, Norwich, United Kingdom	Dave Clements
May 9	*Introduction to Galaxy Workshop*	The Genome Analysis Centre (TGAC), Norwich, United Kingdom	
May 12	*Galaxy Workshop*	University of Edinburgh, Edinburgh, UK	
	Galaxy Project Update	5th Edinburgh Bioinformatics Meeting, University of Edinburgh, Edinburgh, UK	
May 13	*Galaxy Workshop*	Institute of Genetics and Molecular Medicine (IGMM), Edinburgh, UK	
May 12-14	*Short course on RNA-seq and ChIP-seq*	University of Bergen, Bergen, Norway	David Fredman
May 16	*Galaxy Initiation*	Formation en Bioinformatique Plateforme ABiMS, Station Biologique de Roscoff, France	Christophe Caron
May 19	*Initiation au traitement et à l'analyse des données métabolomiques sur la plateforme scientifique web Galaxy IFB-MetaboHUB*	8e Journées Scientifiques du RFMF, Lyon, France	Réseau Français de Métabolomique et Fluxomique

http://wiki.galaxyproject.org/Events

GALAXY COMMUNITY CONFERENCE

BALTIMORE, MD | JUNE 30 - JULY 2, 2014

Training Day: 15 sessions, 2.5 hrs each

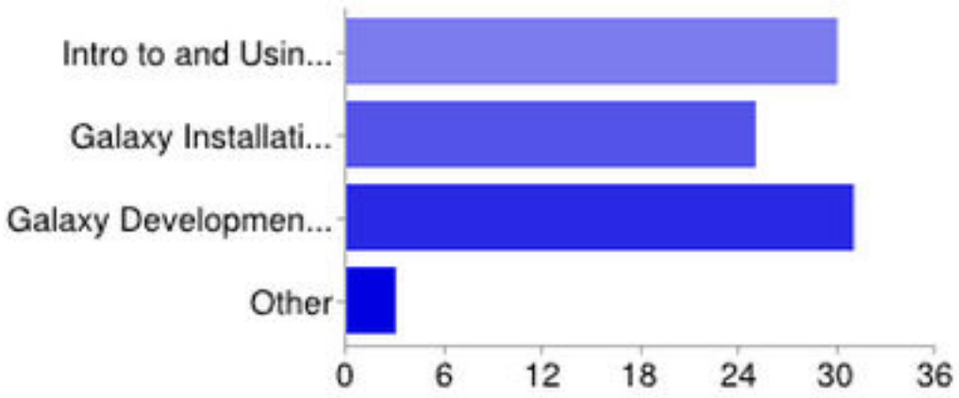
Poster sessions, BoFs

Hackathon

Early registration ends 23 May

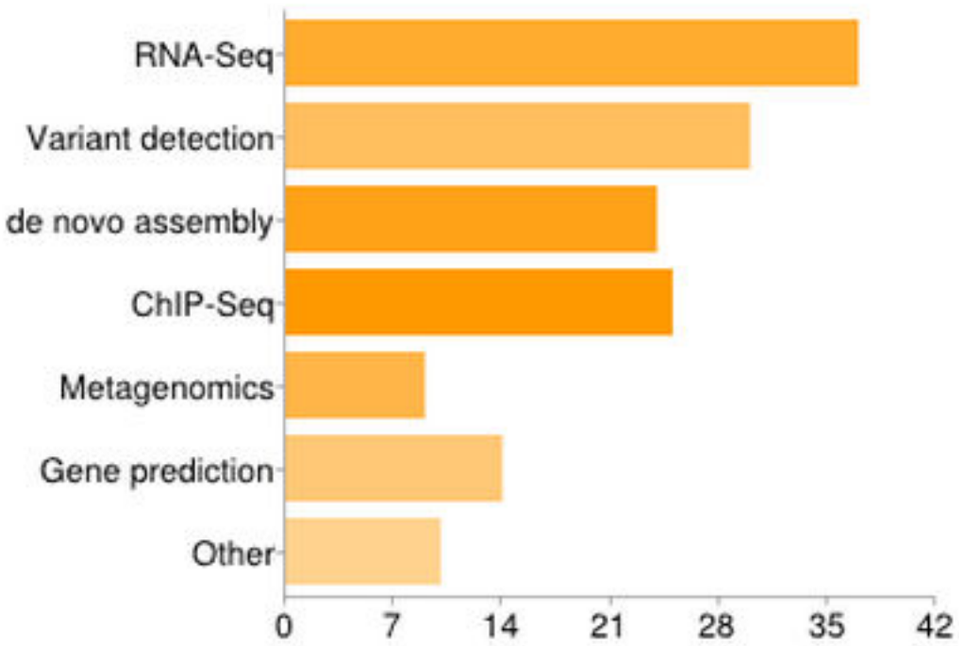
Training: Survey

What type of training are you most interested in?



Intro to and Using Galaxy	**30**	34%
Galaxy Installation and Administraton	**25**	28%
Galaxy Development and Tool Integration	**31**	35%
Other	**3**	3%

What type of data analysis are you interested in?



RNA-Seq	**37**	25%
Variant detection	**30**	20%
de novo assembly	**24**	16%
ChIP-Seq	**25**	17%
Metagenomics	**9**	6%
Gene prediction	**14**	9%
Other	**10**	7%

Training in 2014: Scalability

Create a Training Network
Make it easy to find and use workshop materials created by anyone.
Training AMIs, VMs
Look seriously at MOOCs and Screencasts

Project

Core Team is at 18 people.

1 in Croatia
1 in Australia
10 with Penn State University
5 with Johns Hopkins University
1 with George Washington University

The Galaxy Team

Enis Afgan

Dannon Baker

Dan Blankenberg

Dave Bouvier

Marten Cech

John Chilton

Dave Clements

Nate Coraor

Carl Eberhard

Dorine Francheteau

Jeremy Goecks

Sam Guerler

Jen Jackson

Ross Lazarus

Anton Nekrutenko

Nick Stoler

James Taylor

Greg von Kuster

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

Acknowledgements

Edinburgh

Alastair Kerr
Graeme Grimes
Mick Watson
Shaun Webb
James Prendergast
Bert Overduin

Norwich

Matt Drew
Vicky Schneider-Gricar

edinburgh genomics
design execution analysis

igmm
INSTITUTE OF GENETICS & MOLECULAR MEDICINE

ROSLIN

Wellcome Trust Centre for Cell Biology WTCCB

Edinburgh Bioinformatics

TGAC
The Genome Analysis Centre™

THE UNIVERSITY of EDINBURGH

Thank You!

Events: 2013

200+ Talks, workshops, tutorials, ...

90+ Events

80% With archived slides, video, exercises, ...

75% Presented by the Galaxy Community

GCC2013

Added Poster sessions
Longer training sessions (2 hrs)
Breakout reformed into BoFs
2nd year with sponsors
Had a pub onsite!

GCC attendance over time