

# Galaxy

## Accessible and Reproducible Data Analysis for Bench Scientists

---

Institute of Ecology and Evolution  
University of Oregon  
October 21, 2014

Dave Clements  
Johns Hopkins University

THE INSTITUTE OF ECOLOGY AND EVOLUTION



UNIVERSITY OF OREGON



Galaxy is an open-source, web-based, data integration and analysis platform for life science research. Galaxy enables bench scientists to create, share, and publish sophisticated, reproducible bioinformatic analyses without requiring researchers to learn command line interfaces, or Unix system management skills. Galaxy can be accessed through the project's public server, or on one of the over 60 publicly accessible Galaxy servers. Galaxy can also be installed locally, and on cloud infrastructures.

This talk will introduce the Galaxy platform and discuss the project's recent work and plans going forward. Time allowing, there will also be a brief demonstration.

Galaxy is ...

Galaxy is a ... web-based, data integration and analysis platform for life science research scientists to create, share, and publish sophisticated, reproducible bioinformatics analyses without requiring researchers to learn command line interfaces, or Unix system management skills. Galaxy can be accessed through the project's public server, or on one of the over 60 publicly accessible Galaxy servers. Galaxy can also be installed locally, and on cloud infrastructures.

This talk will introduce the Galaxy platform ...

there will also be a brief demonstration.

Galaxy is ...

...

Galaxy enables bench scientists  
to create ... bioinformatic  
analyses ...  
... there will also be a brief  
demonstration.

This talk will introduce the Galaxy platform and discuss the  
project's recent work and plans going forward. —————  
there will also be a brief demonstration.

# Basic Analysis

Which exons have most overlapping  
repeats in 3 spine stickelback, chromosome  
XXI?

(~ <http://usegalaxy.org/galaxy101> )

Galaxy is

analysis platform for life science research. Galaxy enables bench scientists to create,

**reproducible**

researchers to learn command line interfaces, or Unix system management skills. Galaxy can be accessed through the project's public server, or one of the over 60 publicly accessible Galaxy servers. Galaxy can also be installed locally, and on cloud infrastructures.

...reproducible...

This talk will introduce the Galaxy platform and discuss the project's recent work and plans going forward. —————  
there will also be a brief demonstration.

# Galaxy is ...

...

platform for life science research. Galaxy enables bench scientists to create,

bioinformatic analyses

command line interfaces, or Unix system management skills.

Galaxy can be accessed through the project's public server, or

on one of the over 60 publicly accessible Galaxy servers.

Galaxy can also be installed locally, and on cloud infrastructures.

**... share, and publish ...**

This talk will introduce the Galaxy platform and discuss the

project's recent work and plans going forward. —————

there will also be a brief demonstration.

# Galaxy is ...

...

platform for life science research. Galaxy enables bench scientists to create, share, and publish

bioinformatic analyses

command line interfaces, or Unix system management skills.

Galaxy can be accessed through the project's public server, or on one of the over 50 publicly accessible Galaxy servers.

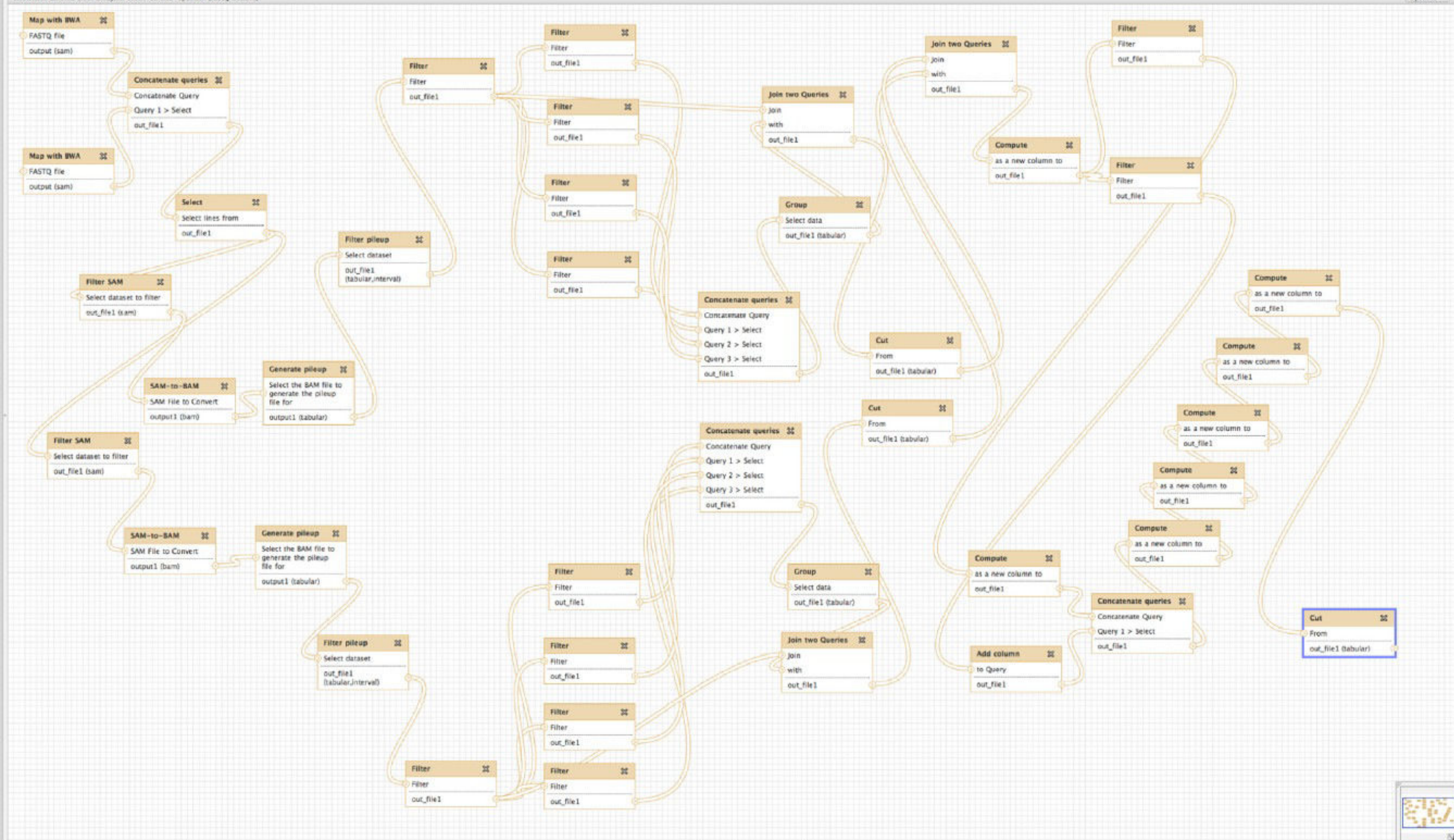
**... sophisticated ...**

Galaxy can also be installed locally, and on cloud infrastructures.

This talk will introduce the Galaxy platform and discuss the project's recent work and plans going forward. —————

there will also be a brief demonstration.





Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study, Goto *et al. Genome Biology* 2011, 12:R59  
<http://genomebiology.com/2011/12/6/R59>

# Galaxy is ...

...

platform for life science research. Galaxy enables bench scientists to create, share, and publish sophisticated, reproducible

**... without requiring researchers  
command line interfaces, or Unix system management skills.  
to learn command line  
interfaces, or Unix system  
management skills.**

This talk will introduce the Galaxy platform and discuss the project's recent work and plans going forward. —————  
there will also be a brief demonstration.

# Galaxy is ...

...

platform for life science research. Galaxy enables bench scientists to create, share, and publish sophisticated, reproducible bioinformatic analyses using a web-based graphical user interface, or Unix system management skills.

**... discuss the project's recent work and plans going forward.**

Galaxy can be accessed through the project's public server, or on one of the over 60 publicly accessible Galaxy servers. Galaxy can also be installed locally,

**This talk will introduce the Galaxy platform and discuss the project's recent work and plans going forward. —**  
allowing



# Scalability ...



# Scalability

Data generation is cheap and will stay cheap.

Scale & complexity of analysis will continue to grow.

More researchers are running bioinformatics analyses of all scales and complexities.

Galaxy needs to scale to the next few orders of magnitude.



# Semantic Scalability: Dataset Collections

Make Galaxy aware of how datasets are related.

Build workflows that can reason about paired datasets, technical replicates, multiple biological samples, ...

Run tools once on each dataset in the collection.  
Run tools on the collection as a whole.

Support map/reduce paradigm.

Project Versions  
latest

## RTD Search

 Go

Full-text doc search.

## Table Of Contents

Galaxy API Documentation

Background

Quickstart

API Controllers

datasets Module

folder\_contents Modu

folders Module

forms Module

genomes Module

group\_roles Module

group\_users Module

groups Module

histories Module

history\_contents Mod

item\_tags Module

libraries Module

library\_contents Mod

permissions Module

quotas Module

request\_types Module

requests Module

roles Module

samples Module

tools Module

users Module

visualizations Modul

workflows Module

## Previous topic

galaxy Package

## Next topic

controllers Package

## This Page

# Scaling for the Bioinformatician: Galaxy API

## 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, scheduling, and archiving

Scaling up also requires support for bioinformaticians and core staff.

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:

Allows compute-savvy researchers to use scripting and still get the reproducibility, sharing, and publishing advantages of Galaxy.

```
% ./display.py my_key http://localhost:4096/api/histories/8c49be448cfe29bc
Collection Members
-----
#1: /api/histories/8c49be448cfe29bc
name: Unnamed history
#2: /api/histories/8c49be448cfe29bc
name: output test
id: 33b43b4e7093c91f
```

The result is a Collection of the histories for the specified ID (the API key (key)). Click on the details of each history, say #1 above, or the collection

```
% ./display.py my_key http://localhost:4096/api/histories/8c49be448cfe29bc
Member Information
-----
state_details: {'ok': 1, 'failed_metadata': 0, 'upload': 0, 'discarded': 0, 'running': 0, 'setting_metadata': 0, 'error': 0, 'new': 0, 'queued': 0, 'e
state: ok
contents_url: /api/histories/8c49be448cfe29bc/contents
id: 8c49be448cfe29bc
name: Unnamed history
```

This gives detailed information about the specific member in question, in this case the History. To view history contents, do the following:

```
% ./display.py my_key http://localhost:4096/api/histories/8c49be448cfe29bc/contents
Collection Members
-----
#1: /api/histories/8c49be448cfe29bc/contents/6f91353f3eb0fa4a
```

galaxy-dist.readthedocs.org

# Galaxy is ...

...

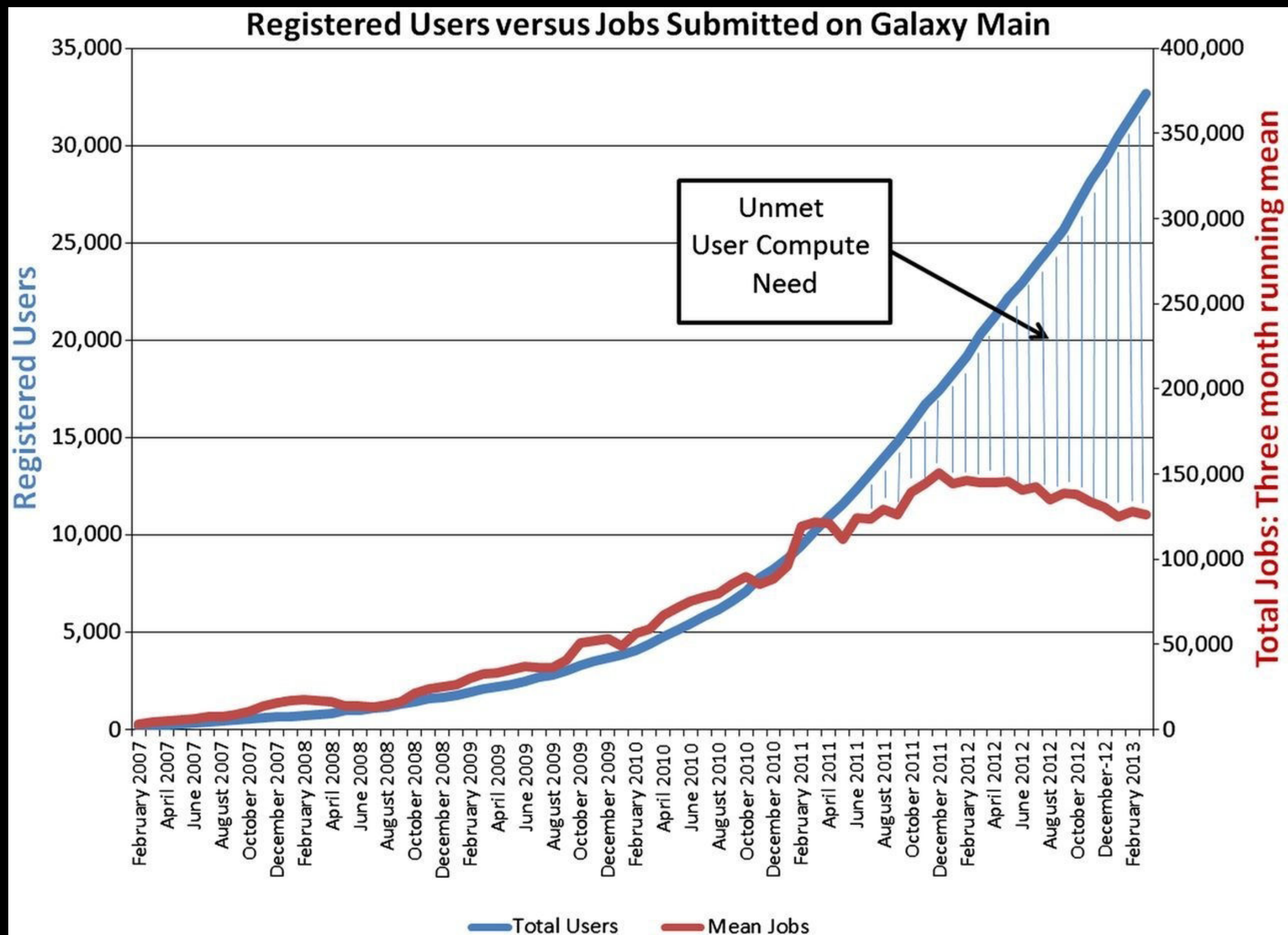
platform for life science research. Galaxy enables bench scientists to create, share, and publish sophisticated, reproducible bioinformatics analyses.

**Galaxy can be accessed through the project's public server ...**

This talk will introduce the Galaxy platform and discuss the project's recent work and plans going forward. —————  
there will also be a brief demonstration.

**[usegalaxy.org](http://usegalaxy.org)**





Leveraging the national cyberinfrastructure for biomedical research

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

# Galaxy is ...

...

platform for life science research. Galaxy enables bench scientists to create, share, and publish sophisticated, reproducible bioinformatic analyses. ... **open-source** ...

command line interfaces, or Unix system management skills.

Galaxy can be accessed through the project's public server, one of the over 50 publicly accessible Galaxy servers.

**Galaxy can also be installed locally ...**

This talk will introduce the Galaxy platform and discuss the project's recent work and plans going forward. —————  
there will also be a brief demonstration.

**getgalaxy.org**

# Galaxy is ...

...

platform for life science research. Galaxy enables bench scientists to create, share, and publish sophisticated, reproducible bioinformatic analyses.

... or on one of the over 60

publicly accessible Galaxy

servers.

This talk will introduce the Galaxy platform and discuss the project's recent work and plans going forward. —————  
there will also be a brief demonstration.

[bit.ly/gxyServers](http://bit.ly/gxyServers)



# Cistrome

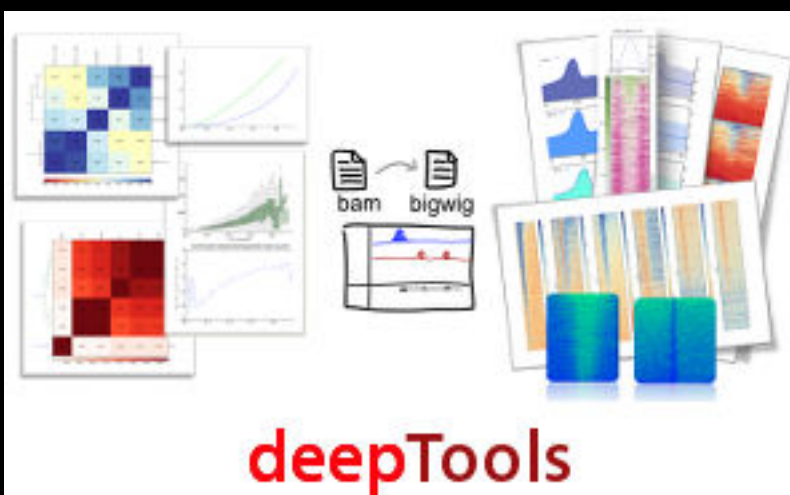


A Galaxy Server  
dedicated to  
ChIP-\* analysis



ballaxy

Powered by the  
Biochemical  
Algorithms  
Library  
Project



deepTools



galaxy.berkeleybop.org

**GWIS: Online exhaustive bivariate GWAS in minutes...**

David R. Krawinkel, Hanyuan Peng, Andrew Krawinkel, Qian Wang, Ben Ouyang, Adam Krawinkel

GWIS (Genome-Wide Interactive Search) is a fast method for detecting bivariate association between genotypes and phenotypes in GWAS data. The algorithm used in GWIS was recently evaluated against conventional methods (1) on 7 Wellcome Trust Case-Control Consortium datasets (2).

Not only was it shown that GWIS methods were faster than all other exhaustive algorithms, but they explicitly search for a well-defined proxy of epistasis. An improvement in association for SNP pairs, over the association for each individual SNP.

**Web Service**

We have now developed a fast online interface to GWIS, based on an instance of the GalaxyProject server (3). Users can upload GWIS datasets in PLINK format (4) for processing with a history of previous computational runs for reproducibility, or using the 3 beta-specific GWIS datasets (5).

The server is free for public use as a demonstration of our methods. It is limited to only a single dataset upload, per exhaustive search analysis (10 Gb). It also can be completed in 10 minutes or a dataset of equal size to the 1000G reference.

For your detailed test, the server returns a complete list of the most significant SNPs, with their p-values, and a list of the most significant SNPs, with their p-values. Up to 1 million SNPs can be tested.

**Timing Data**

Execution times for exhaustive bivariate analysis using the GWIS demo server are reported below. For comparison, we include timing figures reported for PLINK (6), GEM (7), and PLINK (8). The fastest and most widely used alternative methods reported in the literature.

**Processing Pipeline**

Welcome to Cloud-based Image Analysis and Processing Toolbox...

**CloudBased Image Analysis & Processing Toolbox**

More information can be found on the NeCTAR website, and the project blog.

This project is supported in part by NeCTAR, and CSIRO.

**Galaxy / Metabiome Portal**

**The Microbiome Analysis Center**  
Life on a Smaller Scale

Welcome to the Metabiome Portal @ GMU

We have developed the Metabiome Portal, a flexible and customizable framework with the aim of unifying current, image, system, and analysis of microbiome data. The Portal serves as a central hub for data management and analysis, and also provides a platform for data sharing and collaboration.

bit.ly/gxyServers



# Cistrome

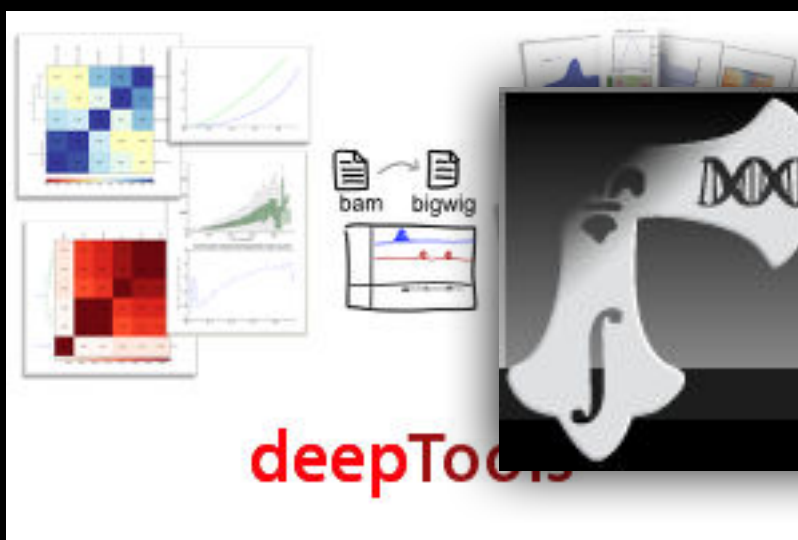


A Galaxy Server  
dedicated to  
ChIP-\* analysis



ballaxy

Powered by the  
Biochemical  
Algorithms  
Library  
Project



deepTools



## The Huttenhower Lab

Department of Biostatistics, Harvard School of Public Health

NEONTOLOGY  
Unifying Biology

berkeleybop.org

**GWIS: Online exhaustive bivariate GWAS in minutes...**

David R. Krawinkel, Hanyu Pan, Andrew Krawinkel, Qian Wang, Ben Ouyang, Adam Krawinkel

GWIS (Genome-Wide Interactive Search) is a fast method for detecting bivariate association between genotypes and phenotypes in GWAS data. The algorithm used in GWIS was recently evaluated against conventional methods (1) on 7 Wellcome Trust Case-Control Consortium datasets (2).

Not only was it shown that GWIS methods were faster than all other exhaustive algorithms, but they explicitly search for a well-defined proxy of epistasis. An improvement in association for SNP pairs, over the association for each individual SNP.

**Web Service**

We have now developed a fast online interface to GWIS, based on an instance of the GalaxyProject server (3). Users can upload GWIS datasets in PLINK format (4) for processing with a history of previous computational runs for reproducibility, or using the 3 beta-specific datasets (5, 6, 7) and a subset of the 1000 Genomes Project (8).

The server is free for public use as a demonstration of the method. It is hosted on a single desktop machine, but exhaustive searches require 4 GB of RAM and can be completed in 10 minutes on a dataset of equal size to the 1000 Genomes Project.

For your detailed test, the server returns a summary list of the most significant SNPs, with their p-values and the associated SNPs. Up to 1 million SNPs can be tested.

**Processing Pipeline**

Welcome to Cloud-based Image Analysis and Processing Toolbox...

**CloudBased Image Analysis & Processing Toolbox**

More information can be found on the NeCTAR website, and the project blog.

This project is supported in part by NeCTAR, and CSIRO.

**Galaxy / Metabiome Portal**

**The Microbiome Analysis Center**  
Life on a Smaller Scale

Welcome to the Metabiome Portal @ GMU

We have developed the Metabiome Portal, a flexible and customizable framework with the aim of unifying current, image, system, and analysis methods of microbiome research. The Portal serves as a central database management system and also hosts analytical tools and visualization tools such as taxonomic analysis, network analysis, and more.

bit.ly/gxyServers

# Galaxy is ...

...

platform for life science research. Galaxy enables bench scientists to create, share, and publish sophisticated, reproducible bioinformatic analyses

command line interfaces, or Unix system management skills.

**... and on cloud infrastructures.**  
Galaxy can be accessed through the project's public servers, on one of the over 60 publicly accessible Galaxy servers. Galaxy can also be installed locally,

This talk will introduce the Galaxy platform and discuss the project's recent work and plans going forward. —————  
there will also be a brief demonstration.

[wiki.galaxyproject.org/Cloud](http://wiki.galaxyproject.org/Cloud)



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

<http://globus.org/>

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



# Scaling the Project: **Community Gatherings**



# GCC 2015

Galaxy Community Conference

6-8th July 2015

The Sainsbury Laboratory  
Norwich, UK

[galaxypproject.org](http://galaxypproject.org)



# Scaling the Project: **Support**

**Tens of thousands of users** leads to a lot of questions.

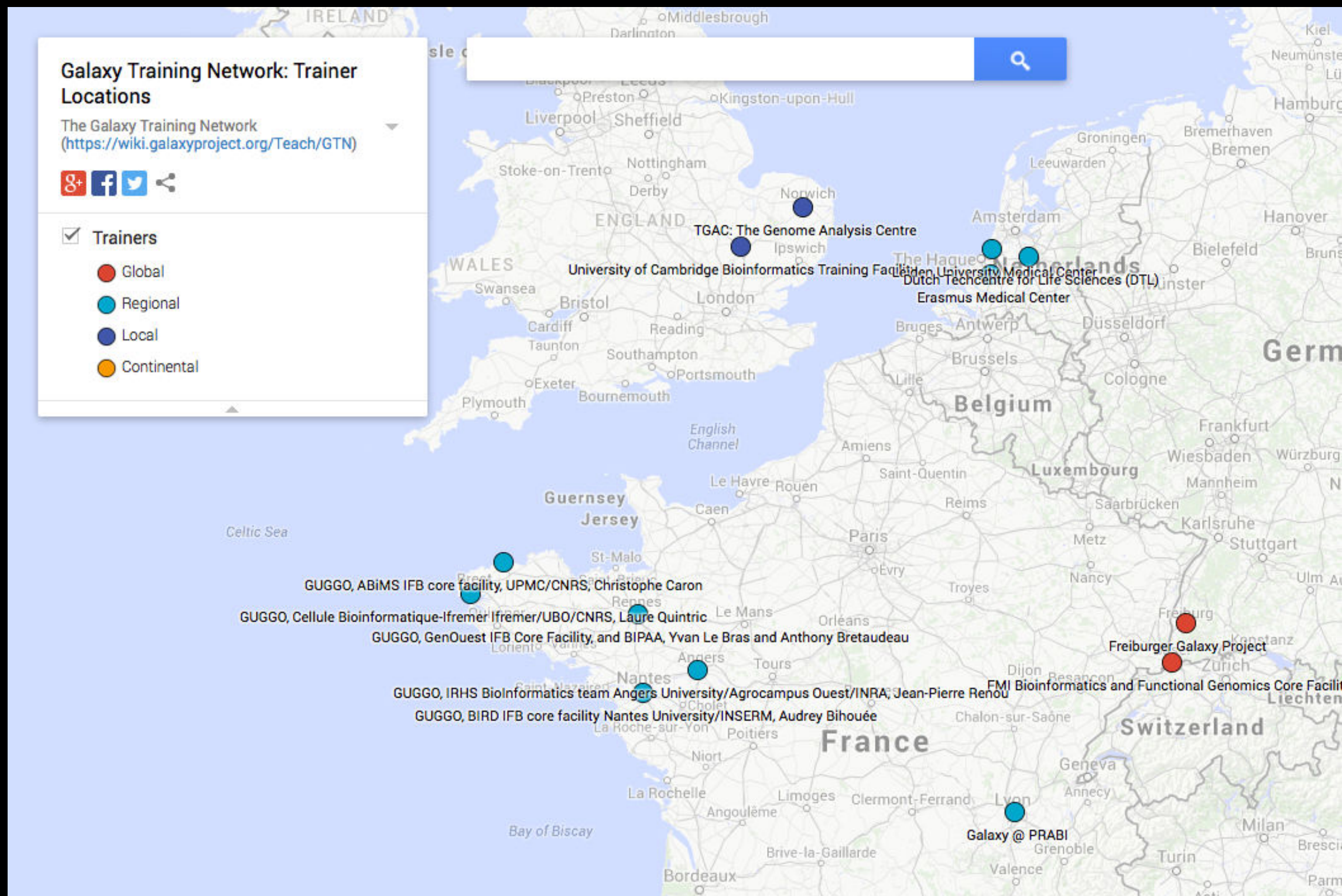
Absolutely have to **encourage community support**.

Project traditionally used mailing list

This year we moved user support to **Galaxy Biostar**, a  
**gamefied online forum**.



# Scaling the Project: Training



Galaxy Training Network launched last week.

[bit.ly/gxygtn](https://bit.ly/gxygtn)

Galaxy is an open-source, web-based, data integration and analysis platform for life science research. Galaxy enables bench scientists to create, share, and publish sophisticated, reproducible bioinformatic analyses without requiring researchers to learn command line interfaces, or Unix system management skills. Galaxy can be accessed through the project's public server, or on one of the over 60 publicly accessible Galaxy servers. Galaxy can also be installed locally, and on cloud infrastructures.

This talk will introduce the Galaxy platform and discuss the project's recent work and plans going forward. Time allowing, there will also be a brief demonstration.



# The Galaxy Team



Enis Afgan



Dannon Baker



Dan Blankenberg



Dave Bouvier



Marten Cech



John Chilton



Dave Clements



Nate Coraor



Carl Eberhard



Jeremy Goecks



Sam Guerler



Jen Jackson



Ross Lazarus



Anton Nekrutenko



Nick Stoler



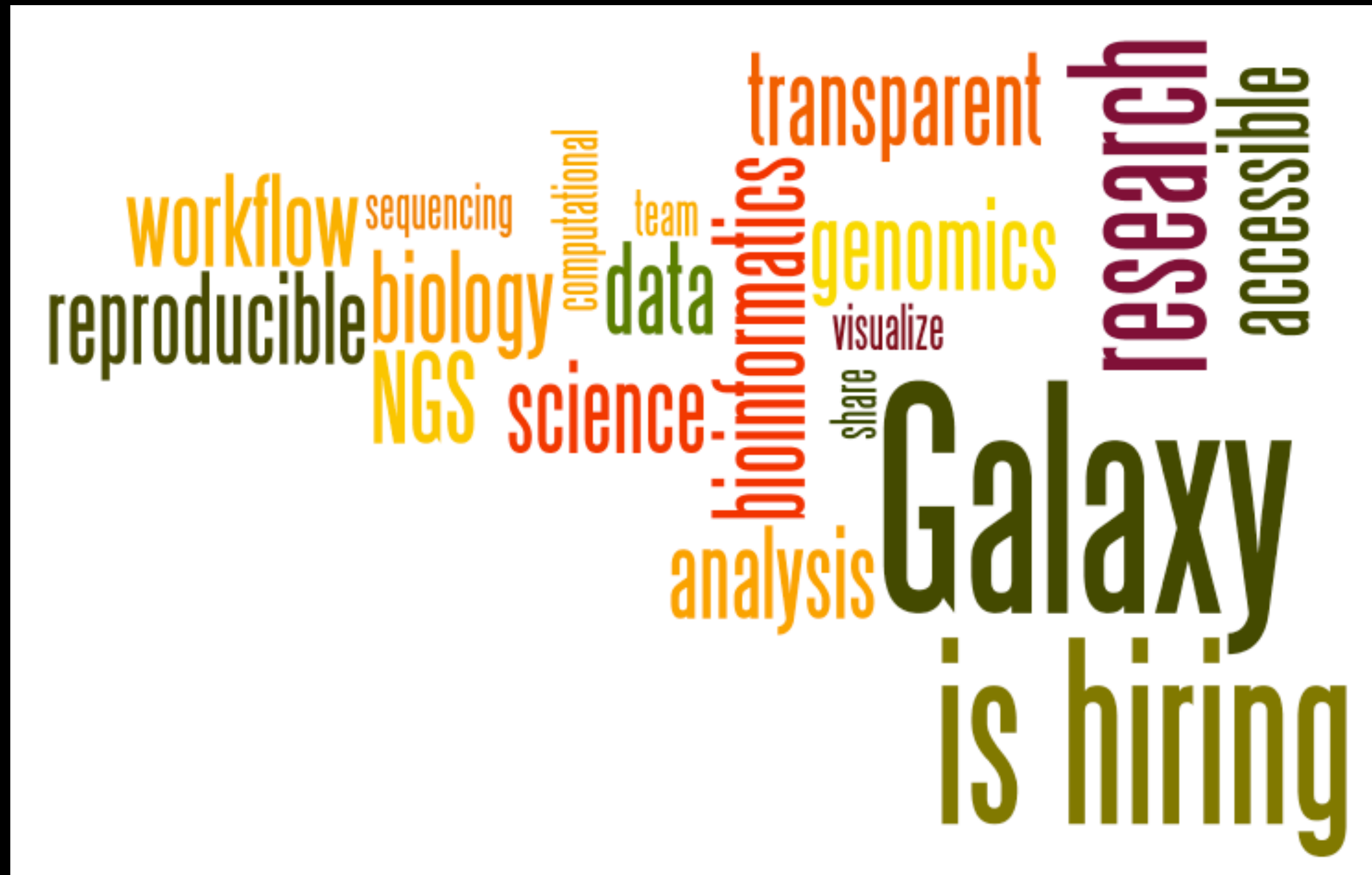
James Taylor



Nitesh Turaga

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

Galaxy is hiring post-docs and software engineers



Please help.

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

Also Thanks To

THE INSTITUTE OF ECOLOGY AND EVOLUTION

Heather Archer  
Michelle Wood  
John Conery

Patrick Phillips  
Bill Cresko



Doug Toomey  
Emilie Hooft



# Thanks



**Dave Clements**

**Galaxy Project**

**Johns Hopkins University**

**[clements@galaxyproject.org](mailto:clements@galaxyproject.org)**