Galaxy

GMOD Workshop
Plant and Animal Genome XX
18 January 2012

Dave Clements, Emory University

http://galaxyproject.org/
http://gmod.org

Agenda

Introduction

Worked example

Deployment Options

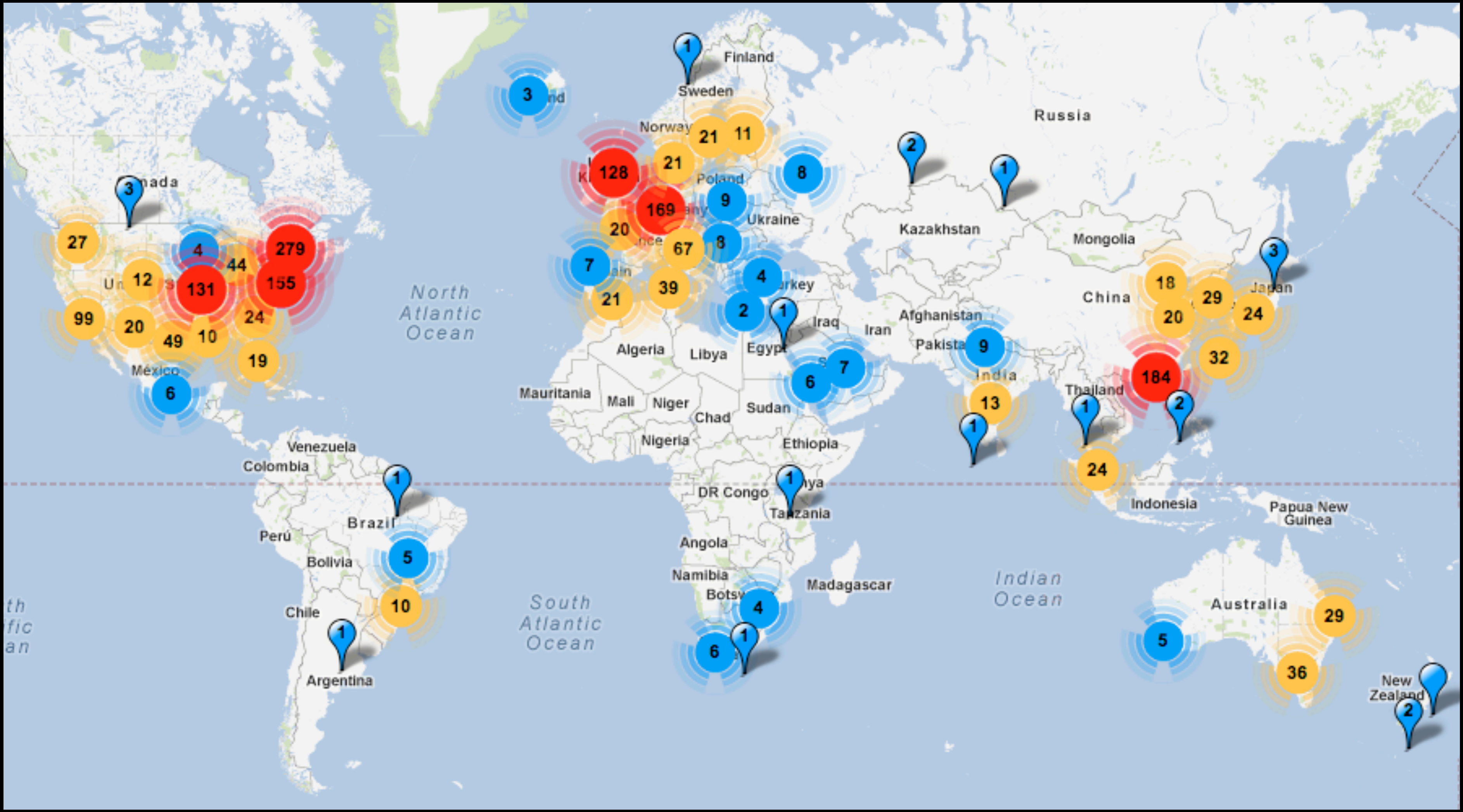
Community

Goals for this workshop

1. Introduce the Galaxy platform.
2. Demonstrate how to:
    - Load and integrate data from popular online resources
    - Perform bioinformatics analysis with Galaxy
    - Save, share, describe and publish your analysis and generated datasets

This workshop will not cover details of how the tools are implemented or new algorithm designs or which assembler or mapper or ... is best for you.

The Motivation Slide



Next Generation Genomics: World Map of High-throughput Sequencers
Nick Loman, James Hadfield

http://pathogenomics.bham.ac.uk/hts/

What is Galaxy?

- **A free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- An **analysis and data integration** tool
- **Open source software** that makes integrating your own tools and data and customizing for your own site simple
- A part of **GMOD**

http://galaxyproject.org

Galaxy URLs to Remember

http://galaxyproject.org

http://usegalaxy.org

http://getgalaxy.org

Agenda

Introduction

Worked example

Deployment Options

Community

Galaxy aims to

Enable
accessible,
reproducible,
and
transparent
computational biomedical research.

Demo: Accessibility

On pig chromosome 18,
which coding exons have the most
repeats in them?

http://usegalaxy.org

Galaxy: A Rough Plan

- Get some data
  - Coding exons on chromosome 18
  - Repeats on chromosome 18
- Mess with it
  - Identify which exons have repeats
  - Count repeats per exon
  - Save, download, visualize, ... exons with most repeats.

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

Galaxy aims to

Enable

accessible,

reproducible,

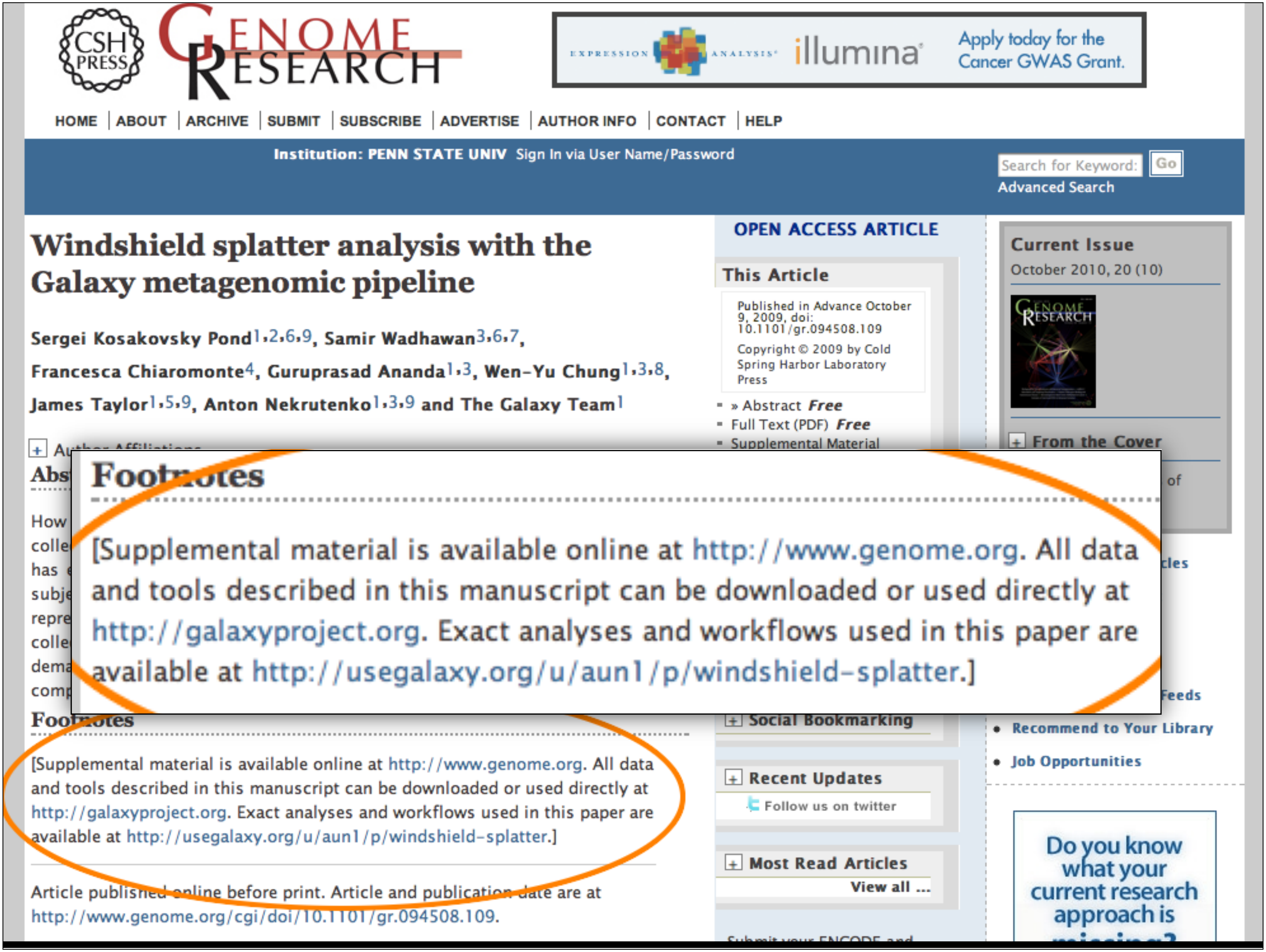
and

transparent

computational biomedical research.

Demo: Reproducibility and Transparency

http://usegalaxy.org



Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond[1,2,6,9], Samir Wadhawan[3,6,7], Francesca Chiaromonte[4], Guruprasad Ananda[1,3], Wen-Yu Chung[1,3,8], James Taylor[1,5,9], Anton Nekrutenko[1,3,9] and The Galaxy Team[1]

Published in Advance October 9, 2009, doi: 10.1101/gr.094508.109

Footnotes

[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.]

Footnotes

[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.]

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.094508.109.

Agenda

Introduction

Worked Example

Deployment Options

Community

Galaxy main site
http://usegalaxy.org

- Public web site, **anybody can use**
- Hundreds of tools
- **Persistent**
- ~500 new users per month, ~100 TB of user data, ~135,000 analysis jobs per month

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 will continue to be maintained and enhanced, but with use limits and storage quotas

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

Scaling Galaxy

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

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
- ***Just download and run, completely self-contained!*** *

* Some assembly required. †

† But not much. ‡

‡ And help is on the way.

Public Galaxy Servers

http://galaxyproject.org/wiki/PublicGalaxyServers

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

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.

CLUSTER RESOURCES
TORQUE

GRID ENGINE

Platform
Computing

DRMAA
Distributed Resource Management
Application API — www.drmaa.org

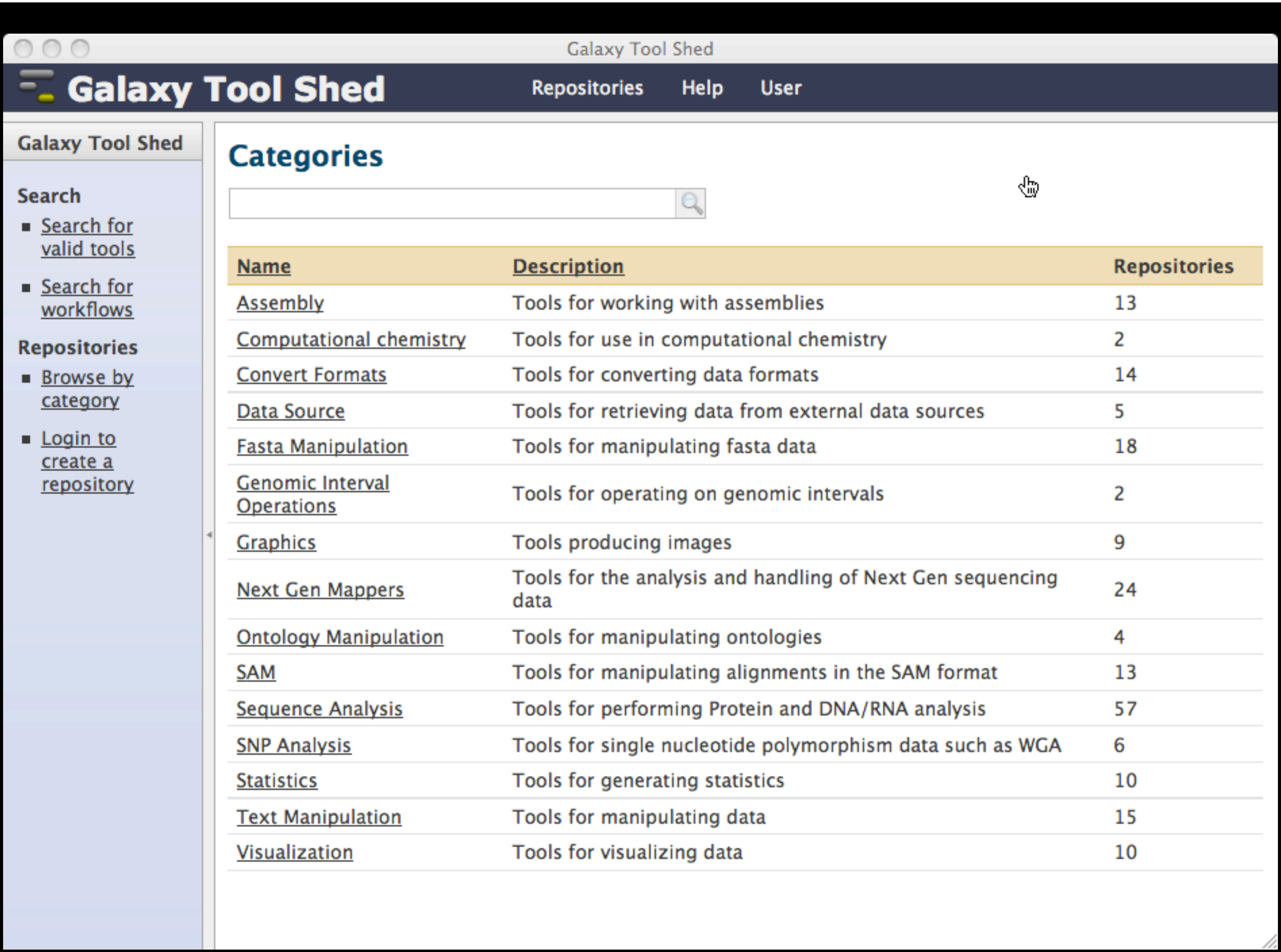
Galaxy CloudMan
http://usegalaxy.org/cloud

- 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.
- Dannon Baker demonstrated this on Monday:
    - In 30 minutes, using only a web browser, Dannon:
    - Setup an elastic compute cluster, fully pre-populated with data and tools, ran some ChIP-Seq analysis, and then shut it down, *and* gave a talk and answered questions.

Galaxy Tool Shed

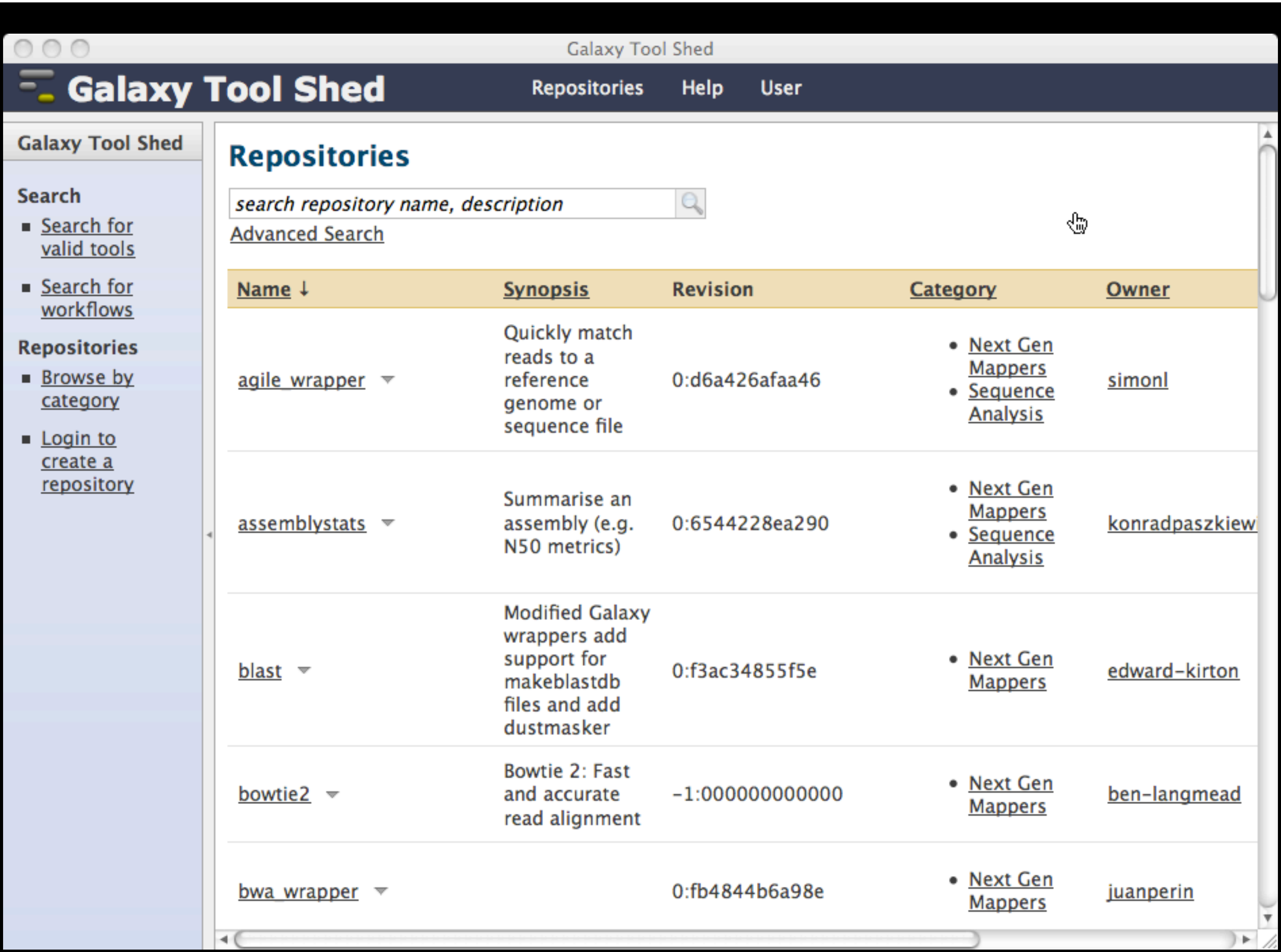
- Allow users to share “suites” containing tools, datatypes, workflows, sample data, and automated installation scripts for tool dependencies
- Integration with Galaxy instances to automate tool installation and updates

http://usegalaxy.org/community



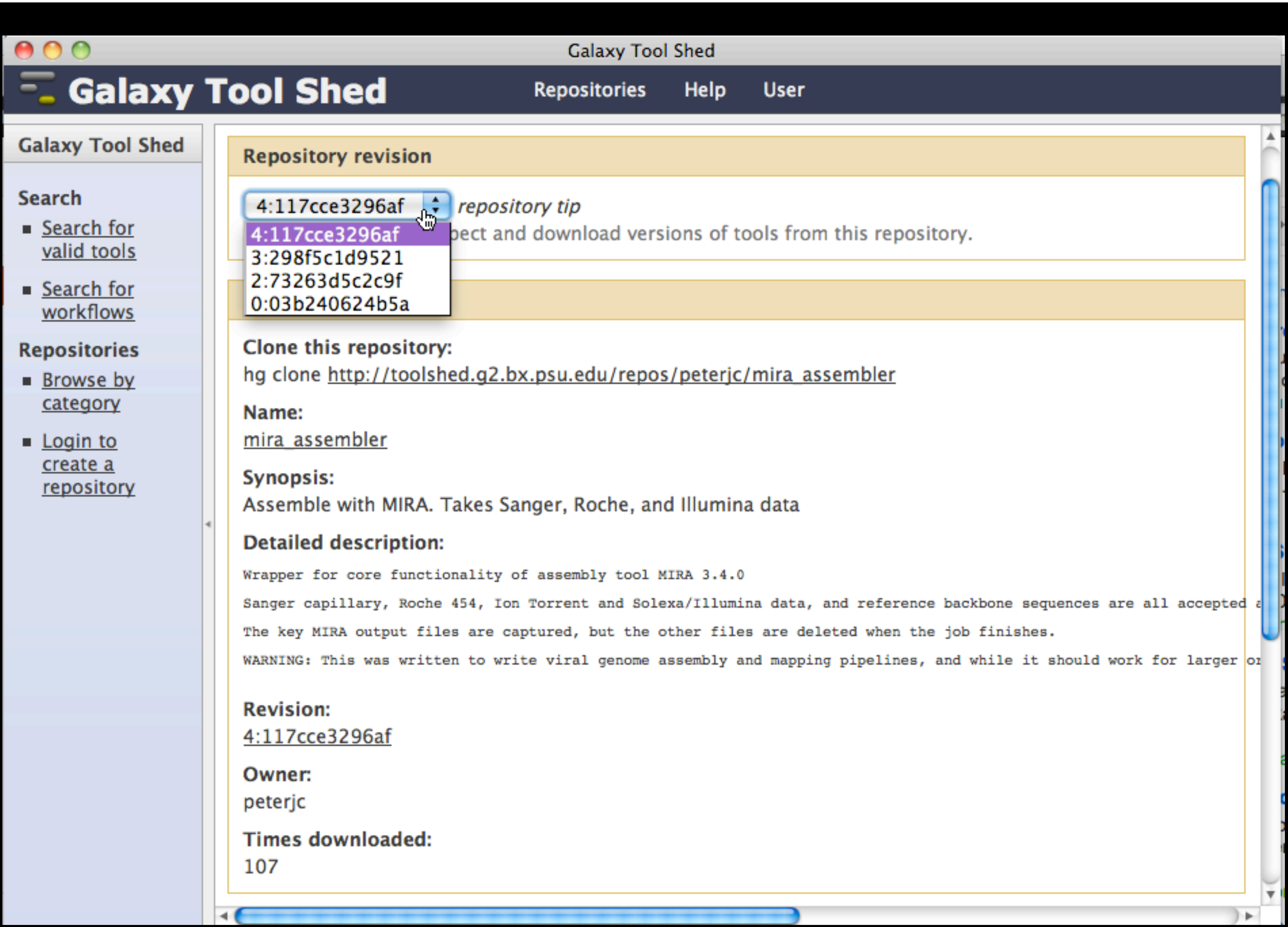
Categories

Name	Description	Repositories
Assembly	Tools for working with assemblies	13
Computational chemistry	Tools for use in computational chemistry	2
Convert Formats	Tools for converting data formats	14
Data Source	Tools for retrieving data from external data sources	5
Fasta Manipulation	Tools for manipulating fasta data	18
Genomic Interval Operations	Tools for operating on genomic intervals	2
Graphics	Tools producing images	9
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	24
Ontology Manipulation	Tools for manipulating ontologies	4
SAM	Tools for manipulating alignments in the SAM format	13
Sequence Analysis	Tools for performing Protein and DNA/RNA analysis	57
SNP Analysis	Tools for single nucleotide polymorphism data such as WGA	6
Statistics	Tools for generating statistics	10
Text Manipulation	Tools for manipulating data	15
Visualization	Tools for visualizing data	10



Repositories

Name ↓	Synopsis	Revision	Category	Owner
agile_wrapper	Quickly match reads to a reference genome or sequence file	0:d6a426afaa46	Next Gen Mappers Sequence Analysis	simonl
assemblystats	Summarise an assembly (e.g. N50 metrics)	0:6544228ea290	Next Gen Mappers Sequence Analysis	konradpaszkiew
blast	Modified Galaxy wrappers add support for makeblastdb files and add dustmasker	0:f3ac34855f5e	Next Gen Mappers	edward-kirton
bowtie2	Bowtie 2: Fast and accurate read alignment	-1:000000000000	Next Gen Mappers	ben-langmead
bwa_wrapper		0:fb4844b6a98e	Next Gen Mappers	juanperin



Galaxy Tool Shed

Search

- Search for valid tools
- Search for workflows

Repositories

- Browse by category
- Login to create a repository

Repository revision

4:117cce3296af *repository tip*

4:117cce3296af
3:298f5c1d9521
2:73263d5c2c9f
0:03b240624b5a

pect and download versions of tools from this repository.

Clone this repository:
hg clone http://toolshed.g2.bx.psu.edu/repos/peterjc/mira_assembler

Name:
mira_assembler

Synopsis:
Assemble with MIRA. Takes Sanger, Roche, and Illumina data

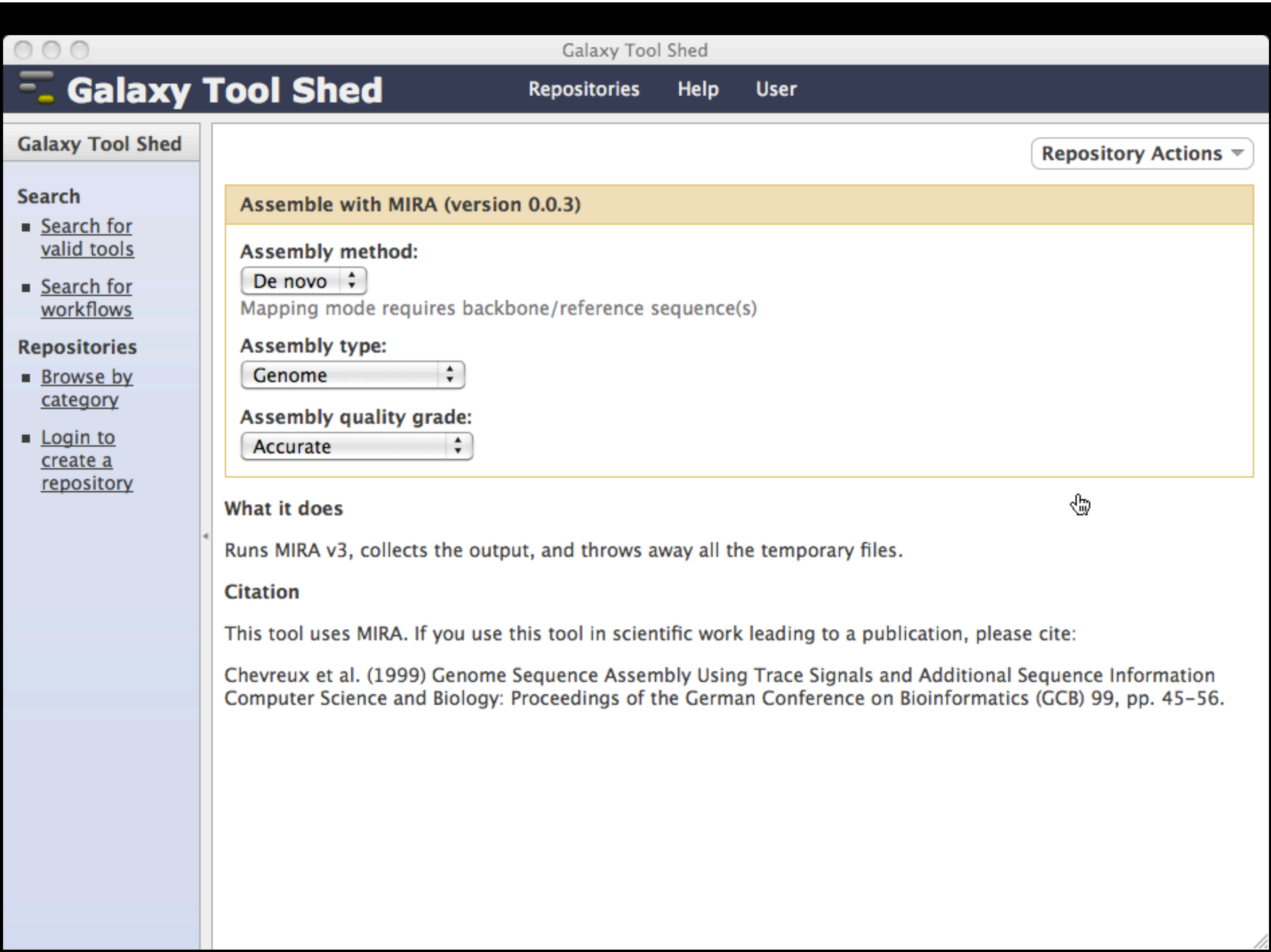
Detailed description:

```
Wrapper for core functionality of assembly tool MIRA 3.4.0

Sanger capillary, Roche 454, Ion Torrent and Solexa/Illumina data, and reference backbone sequences are all accepted a

The key MIRA output files are captured, but the other files are deleted when the job finishes.

WARNING: This was written to write viral genome assembly and mapping pipelines, and while it should work for larger or
```

Revision:
4:117cce3296af

Owner:
peterjc

Times downloaded:
107



Assemble with MIRA (version 0.0.3)

Assembly method: De novo

Mapping mode requires backbone/reference sequence(s)

Assembly type: Genome

Assembly quality grade: Accurate

What it does

Runs MIRA v3, collects the output, and throws away all the temporary files.

Citation

This tool uses MIRA. If you use this tool in scientific work leading to a publication, please cite:

Chevreux et al. (1999) Genome Sequence Assembly Using Trace Signals and Additional Sequence Information Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99, pp. 45-56.

Agenda

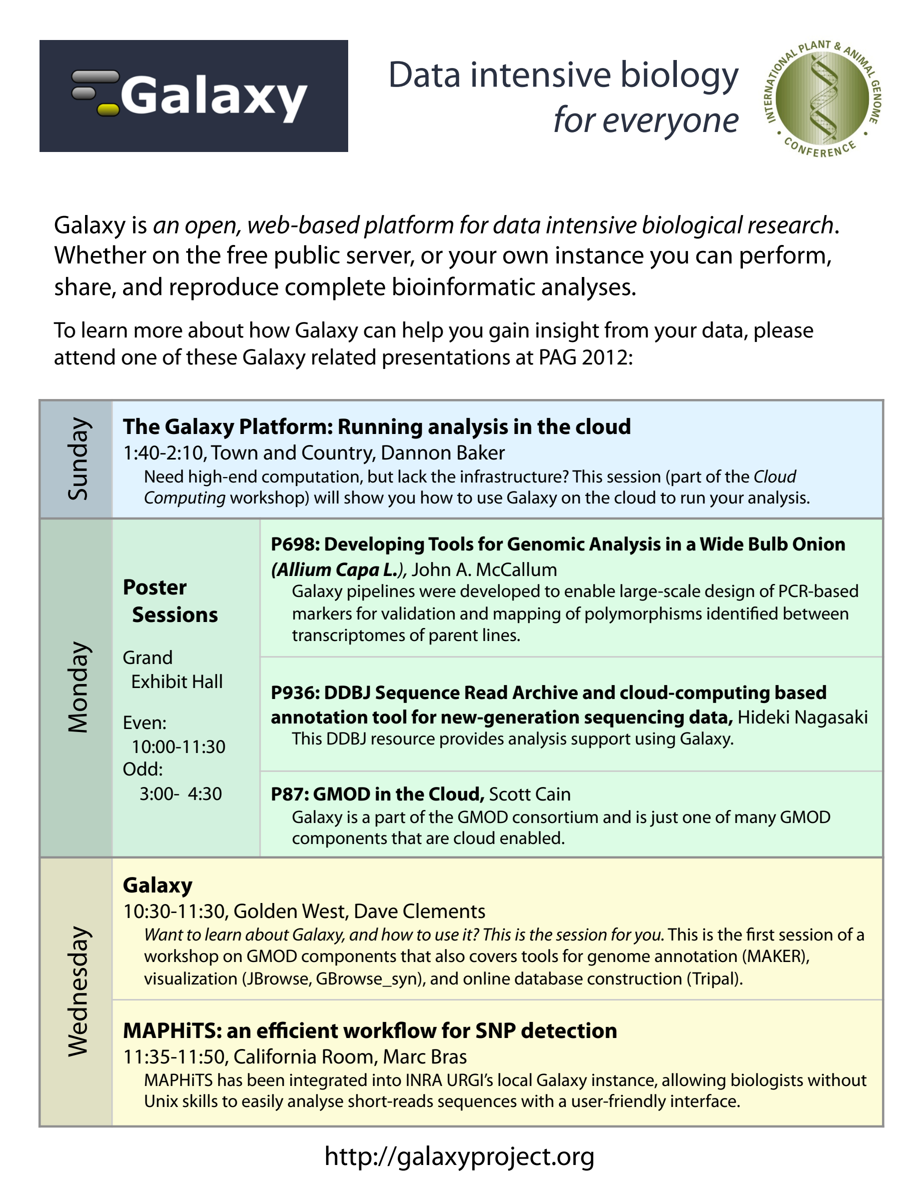
Introduction

Worked Example

Deployment Options

Community

Galaxy Community



Galaxy

Data intensive biology *for everyone*

INTERNATIONAL PLANT & ANIMAL GENOME CONFERENCE

Galaxy is *an open, web-based platform for data intensive biological research*. Whether on the free public server, or your own instance you can perform, share, and reproduce complete bioinformatic analyses.

To learn more about how Galaxy can help you gain insight from your data, please attend one of these Galaxy related presentations at PAG 2012:

Sunday	**The Galaxy Platform: Running analysis in the cloud** 1:40-2:10, Town and Country, Dannon Baker Need high-end computation, but lack the infrastructure? This session (part of the *Cloud Computing* workshop) will show you how to use Galaxy on the cloud to run your analysis.	
Monday	**Poster Sessions** Grand Exhibit Hall Even: 10:00-11:30 Odd: 3:00- 4:30	**P698: Developing Tools for Genomic Analysis in a Wide Bulb Onion *(Allium Capa L.*)**, John A. McCallum Galaxy pipelines were developed to enable large-scale design of PCR-based markers for validation and mapping of polymorphisms identified between transcriptomes of parent lines.
		P936: DDBJ Sequence Read Archive and cloud-computing based annotation tool for new-generation sequencing data, Hideki Nagasaki This DDBJ resource provides analysis support using Galaxy.
		P87: GMOD in the Cloud, Scott Cain Galaxy is a part of the GMOD consortium and is just one of many GMOD components that are cloud enabled.
Wednesday	**Galaxy** 10:30-11:30, Golden West, Dave Clements *Want to learn about Galaxy, and how to use it? This is the session for you.* This is the first session of a workshop on GMOD components that also covers tools for genome annotation (MAKER), visualization (JBrowse, GBrowse_syn), and online database construction (Tripal).	
	MAPHiTS: an efficient workflow for SNP detection 11:35-11:50, California Room, Marc Bras MAPHiTS has been integrated into INRA URGI's local Galaxy instance, allowing biologists without Unix skills to easily analyse short-reads sequences with a user-friendly interface.	

http://galaxyproject.org

Galaxy Community Conference 2012

25-27 July 2012

University of Illinois at Chicago

- Annual Community Meting
- Tool Shed
- Mailing Lists (very active)
- Screencasts
- Events Calendar, News Feed
- Community Wiki
- Local Public Installs

http://galaxyproject.org/wiki

Try it now:
http://UseGalaxy.org

Develop and deploy:
http://GetGalaxy.org

EMORY

PENNSTATE
1855

Dannon Baker

Jeremy Goecks

L

James Taylor

Enis Afgan
(IRB)

Ross Lazarus
(Baker IDI, Harvard)

Dan Blankenberg

Nate Coraor

Jennifer Jackson

Greg von Kuster

Anton Nekrutenko

Supported by the **NHGRI** (HG005542, HG004909, HG005133), **NSF** (DBI-0850103), Penn State University, Emory University, and the Pennsylvania Department of Public Health

http://GalaxyProject.org

Thanks

GMOD

Dr. Scott Cain

PAG Organizers