Galaxy

for high-throughput sequence data analysis

http://usegalaxy.org

The Galaxy Team

Enis Afgan

Guru Ananda

Dannon Baker

Dan Blankenberg

Ramkrishna Chakrabarty

Nate Coraor

Jeremy Goecks

Jennifer Jackson

Greg von Kuster

Kanwei Li

Kelly Vincent

Anton Nekrutenko

James Taylor

Supported by the National Human Genome Research Institute (HG005542, HG004909, HG005133), the National Science Foundation (DBI-0850103), Penn State University, Emory University, and the Pennsylvania Department of Public Health

Are data intensive techniques accessible to researchers?

- For example, high-throughput sequencing:
  - Increasingly availability of instruments, with different strengths, enabling a huge number of high-throughput functional assays
  - However, making use of these techniques requires sophisticated and computationally intensive approaches

Fundamental questions

- When Biology (or any science) becomes dependent on computational methods:
  - How can those methods best be made **accessible** to scientists?
  - How best to facilitate **transparent** communication of those analysis?
  - How best to ensure that analysis are **reproducible**?

A crisis in genomics research:

reproducibility

Key Reproducibility Problems

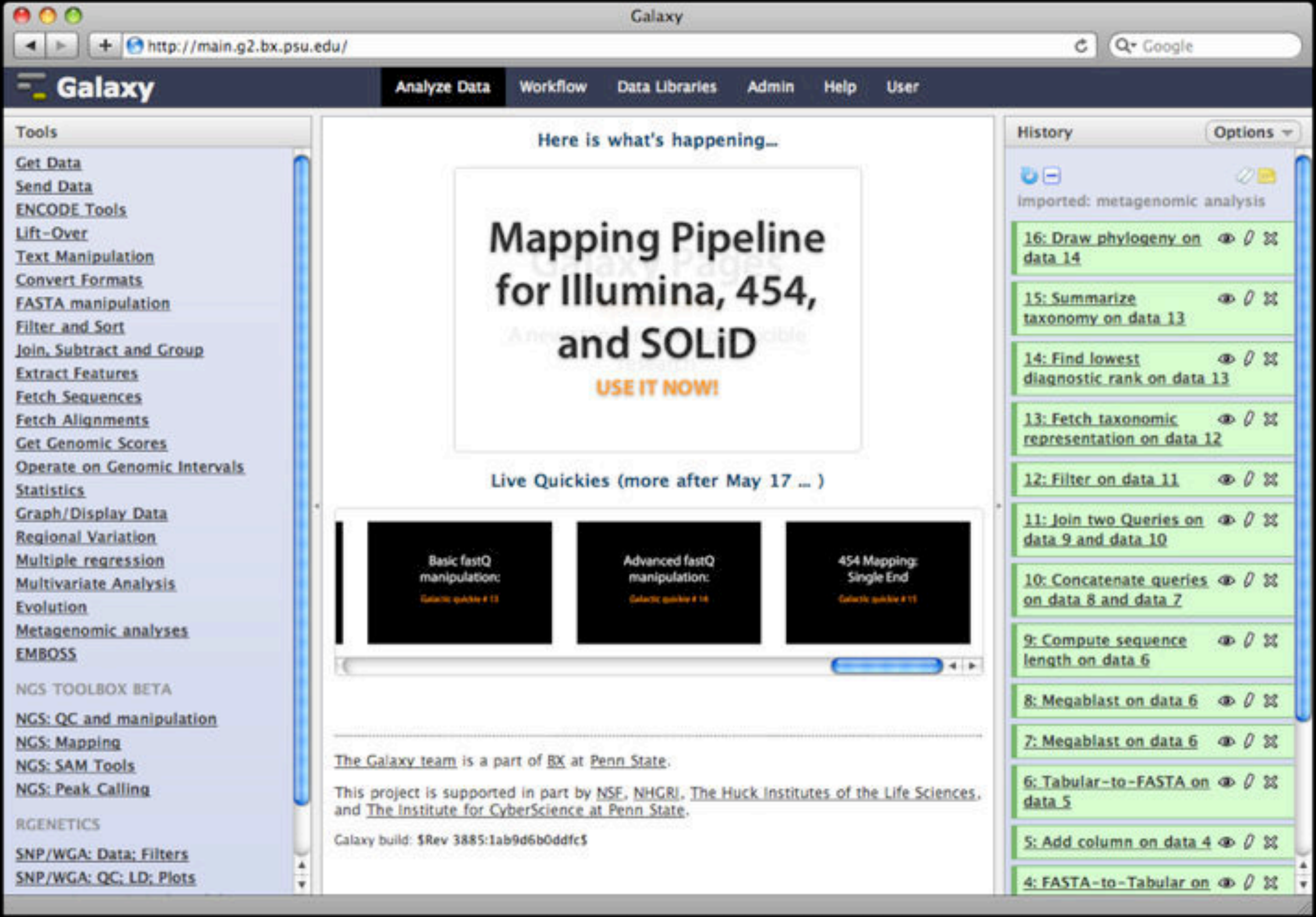
- **Datasets**: not all available, difficult to access
- **Tools**: inaccessible, hard to record details
- **Publication**: results, data, methods separate

Microarray Experiment Reproducibility

- 18 Nat. Genetics microarray gene expression experiments
- Less than 50% reproducible
- Problems
  - missing data (38%)
  - missing software, hardware details (50%)
  - missing method, processing details (66%)

Ioannidis, J.P.A. et al. Repeatability of published microarray gene expression analyses. Nat Genet 41, 149-155 (2009)

Galaxy: accessible analysis system



Here is what's happening...

Mapping Pipeline for Illumina, 454, and SOLiD

USE IT NOW!

Live Quickies (more after May 17 ...)

The Galaxy team is a part of BX at Penn State.

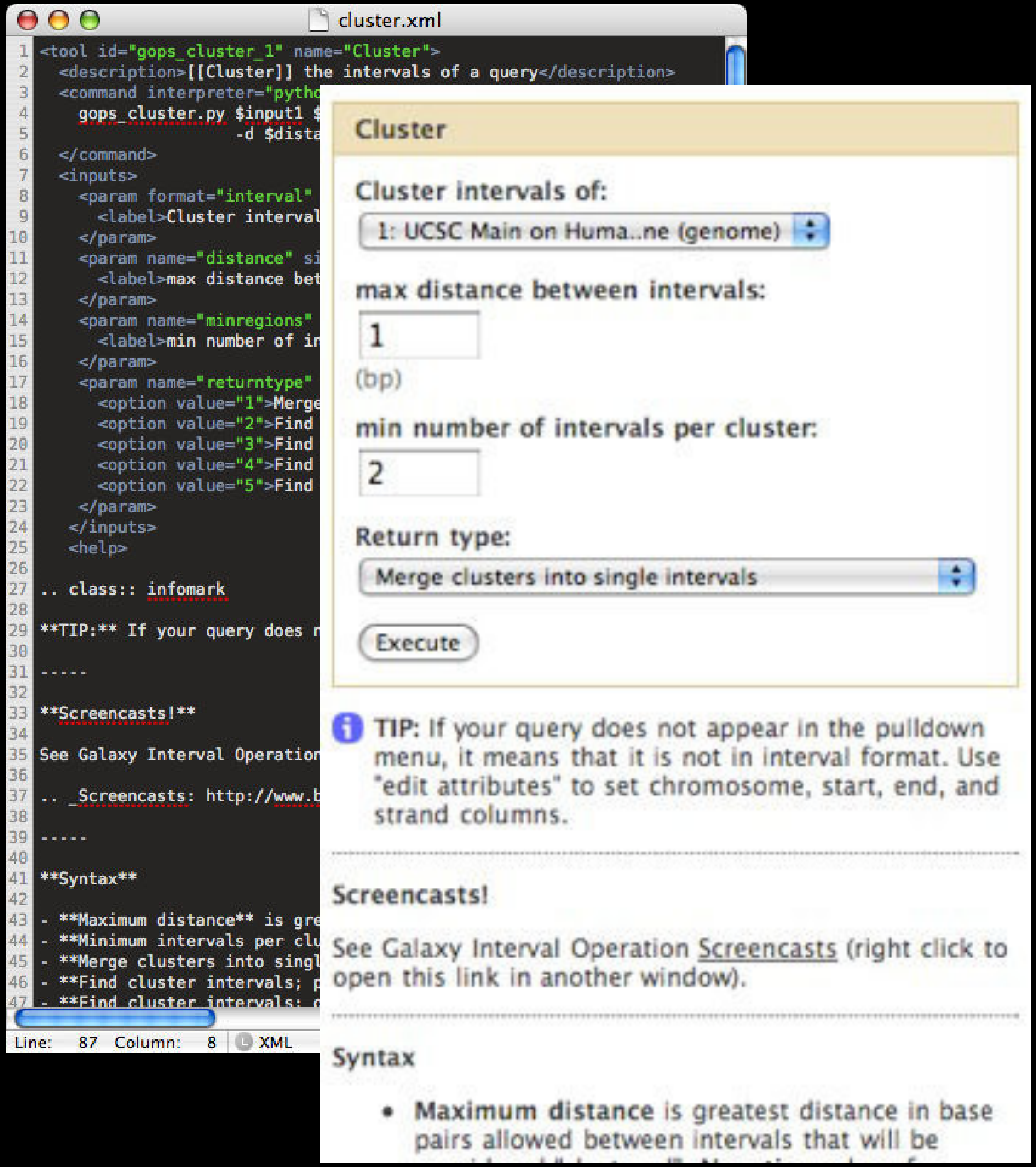
This project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, and The Institute for CyberScience at Penn State.

Galaxy build: $Rev 3885:1ab9d6b0ddfc$

What is Galaxy?

- **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

Integrating existing tools into a uniform framework



cluster.xml

```
<tool id="gops_cluster_1" name="Cluster">
  <description>[[Cluster]] the intervals of a query</description>
  <command interpreter="pytho
    gops_cluster.py $input1 $
                    -d $dista
  </command>
  <inputs>
    <param format="interval"
      <label>Cluster interval
    </param>
    <param name="distance" si
      <label>max distance bet
    </param>
    <param name="minregions"
      <label>min number of ir
    </param>
    <param name="returntype"
      <option value="1">Merge
      <option value="2">Find
      <option value="3">Find
      <option value="4">Find
      <option value="5">Find
    </param>
   </inputs>
   <help>

.. class:: infomark

**TIP:** If your query does r

-----

**Screencasts!**

See Galaxy Interval Operation

.. _Screencasts: http://www.b

-----

**Syntax**

- **Maximum distance** is gre
- **Minimum intervals per clu
- **Merge clusters into singl
- **Find cluster intervals; p
- **Find cluster intervals; c
```

Line: 87 Column: 8 XML

Cluster

Cluster intervals of: 1: UCSC Main on Huma..ne (genome)

max distance between intervals: 1

(bp)

min number of intervals per cluster: 2

Return type: Merge clusters into single intervals

TIP: If your query does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.

Screencasts!

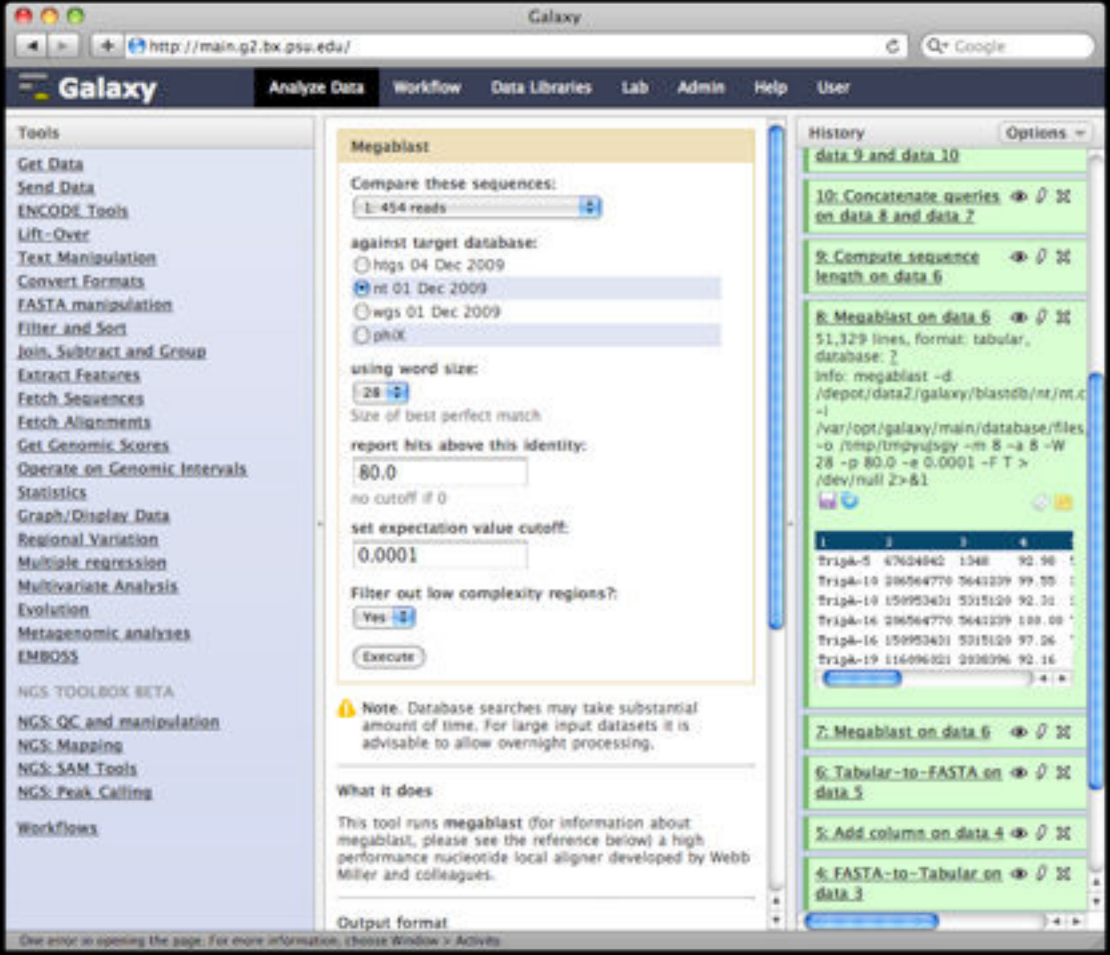
See Galaxy Interval Operation Screencasts (right click to open this link in another window).

Syntax

- **Maximum distance** is greatest distance in base pairs allowed between intervals that will be

- Defined in terms of an abstract interface (inputs and outputs)
  - In practice, mostly command line tools, a declarative XML description of the interface, how to generate a command line
- Designed to be as easy as possible for tool authors, while still allowing rigorous reasoning

Galaxy analysis interface



Megablast

Compare these sequences: 1: 454 reads

against target database:
○ hgs 04 Dec 2009
● nt 01 Dec 2009
○ wgs 01 Dec 2009
○ phiX

using word size: 28
Size of best perfect match

report hits above this identity: 80.0
no cutoff if 0

set expectation value cutoff: 0.0001

Filter out low complexity regions? Yes

Note. Database searches may take substantial amount of time. For large input datasets it is advisable to allow overnight processing.

What it does

This tool runs **megablast** (for information about megablast, please see the reference below) a high performance nucleotide local aligner developed by Webb Miller and colleagues.

Output format

data 9 and data 10

10: Concatenate queries on data 8 and data 7

9: Compute sequence length on data 6

8: Megablast on data 6

51,329 lines, format: tabular, database: ?

Info: megablast -d /depot/data2/galaxy/blastdb/nt/nt.c -i /var/opt/galaxy/main/database/files -o /tmp/tmpyujsgy -m 8 -a 8 -W 28 -p 80.0 -e 0.0001 -F T > /dev/null 2>&1

```
1        2         3        4
frip4-5  47624842  1348     92.98
frip4-10 206564770 5641239  99.55
frip4-10 150953431 5315120  92.31
frip4-16 206564770 5641239  100.00
frip4-16 150953431 5315120  97.26
frip4-19 116096321 2936396  92.16
```

7: Megablast on data 6

6: Tabular-to-FASTA on data 5

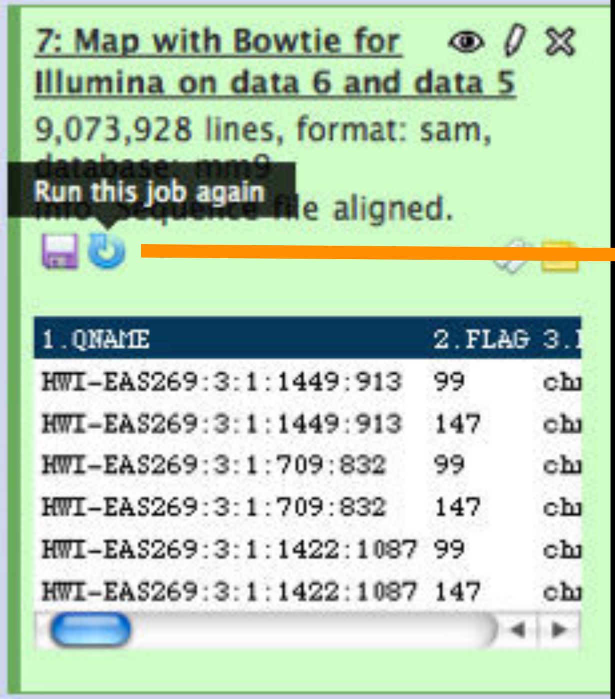
5: Add column on data 4

4: FASTA-to-Tabular on data 3

One error in opening the page. For more information, choose Window > Activity

- Consistent tool user interfaces automatically generated
- History system facilitates and tracks multistep analyses

Automatically tracks every step of every analysis

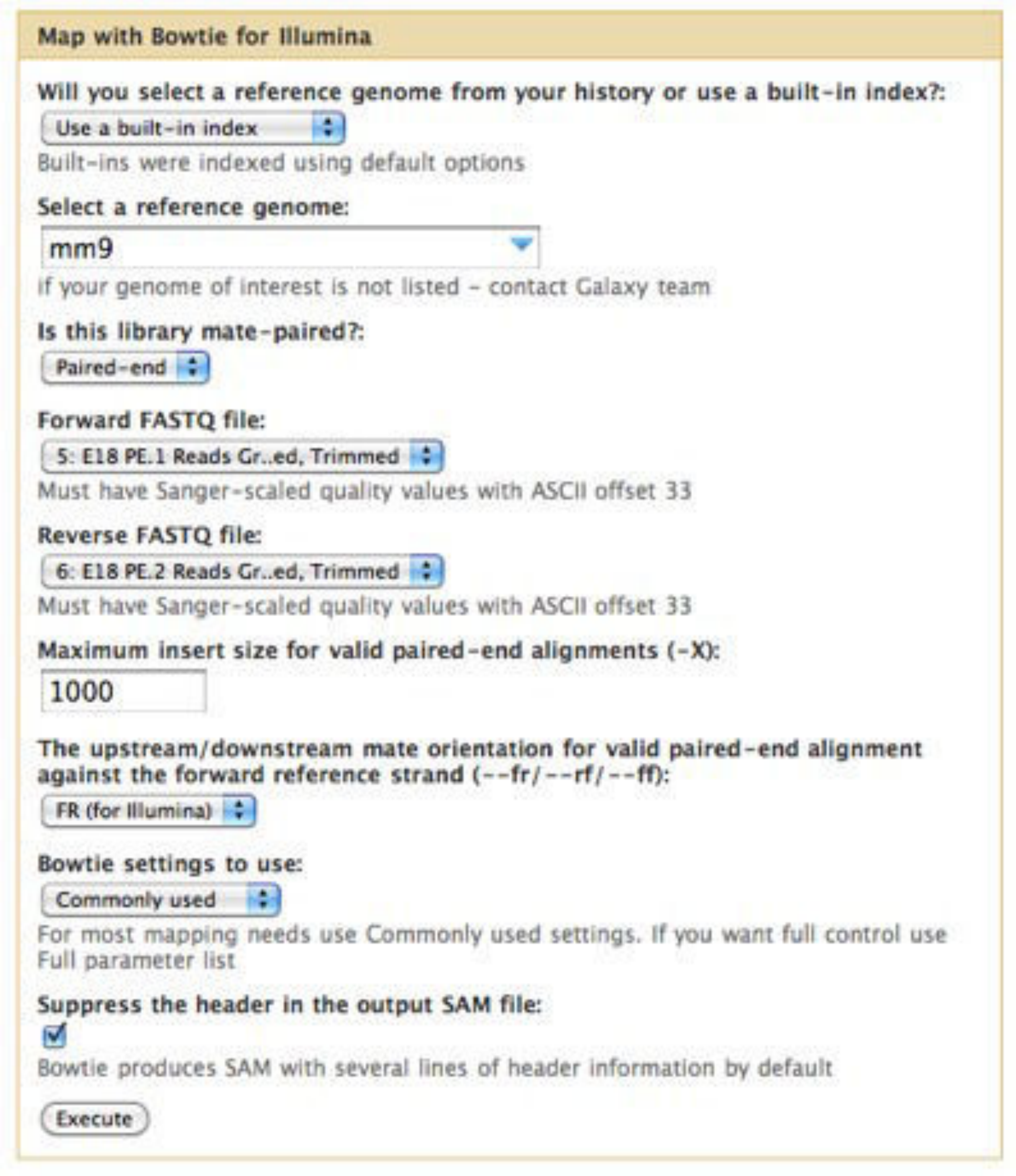


7: Map with Bowtie for Illumina on data 6 and data 5

9,073,928 lines, format: sam, database: mm9

info: sequence file aligned.

1.QNAME	2.FLAG	3.I
HWI-EAS269:3:1:1449:913	99	ch
HWI-EAS269:3:1:1449:913	147	ch
HWI-EAS269:3:1:709:832	99	ch
HWI-EAS269:3:1:709:832	147	ch
HWI-EAS269:3:1:1422:1087	99	ch
HWI-EAS269:3:1:1422:1087	147	ch



Map with Bowtie for Illumina

Will you select a reference genome from your history or use a built-in index?: Use a built-in index

Built-ins were indexed using default options

Select a reference genome: mm9

If your genome of interest is not listed - contact Galaxy team

Is this library mate-paired?: Paired-end

Forward FASTQ file: 5: E18 PE.1 Reads Gr..ed, Trimmed

Must have Sanger-scaled quality values with ASCII offset 33

Reverse FASTQ file: 6: E18 PE.2 Reads Gr..ed, Trimmed

Must have Sanger-scaled quality values with ASCII offset 33

Maximum insert size for valid paired-end alignments (-X): 1000

The upstream/downstream mate orientation for valid paired-end alignment against the forward reference strand (--fr/--rf/--ff): FR (for Illumina)

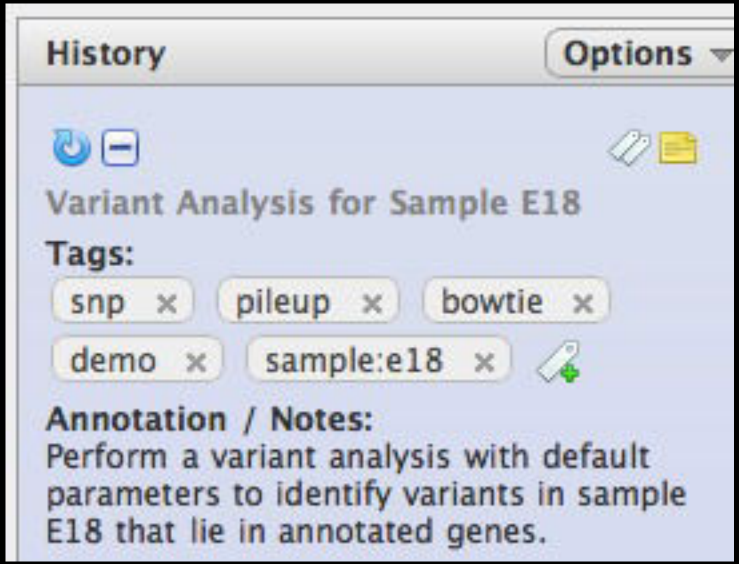
Bowtie settings to use: Commonly used

For most mapping needs use Commonly used settings. If you want full control use Full parameter list

Suppress the header in the output SAM file: ☑

Bowtie produces SAM with several lines of header information by default

As well as user-generated metadata and annotation...

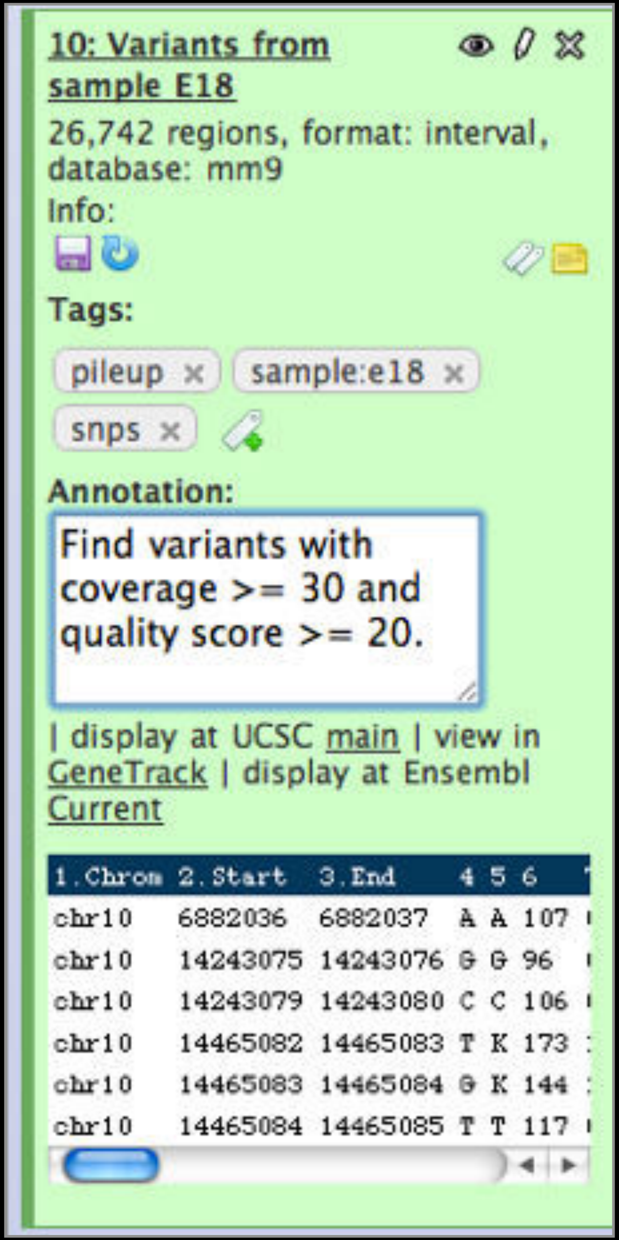


Variant Analysis for Sample E18

Tags:

snp × pileup × bowtie × demo × sample:e18 ×

Annotation / Notes:
Perform a variant analysis with default parameters to identify variants in sample E18 that lie in annotated genes.



10: Variants from sample E18

26,742 regions, format: interval, database: mm9

Info:

Tags:

pileup × sample:e18 × snps ×

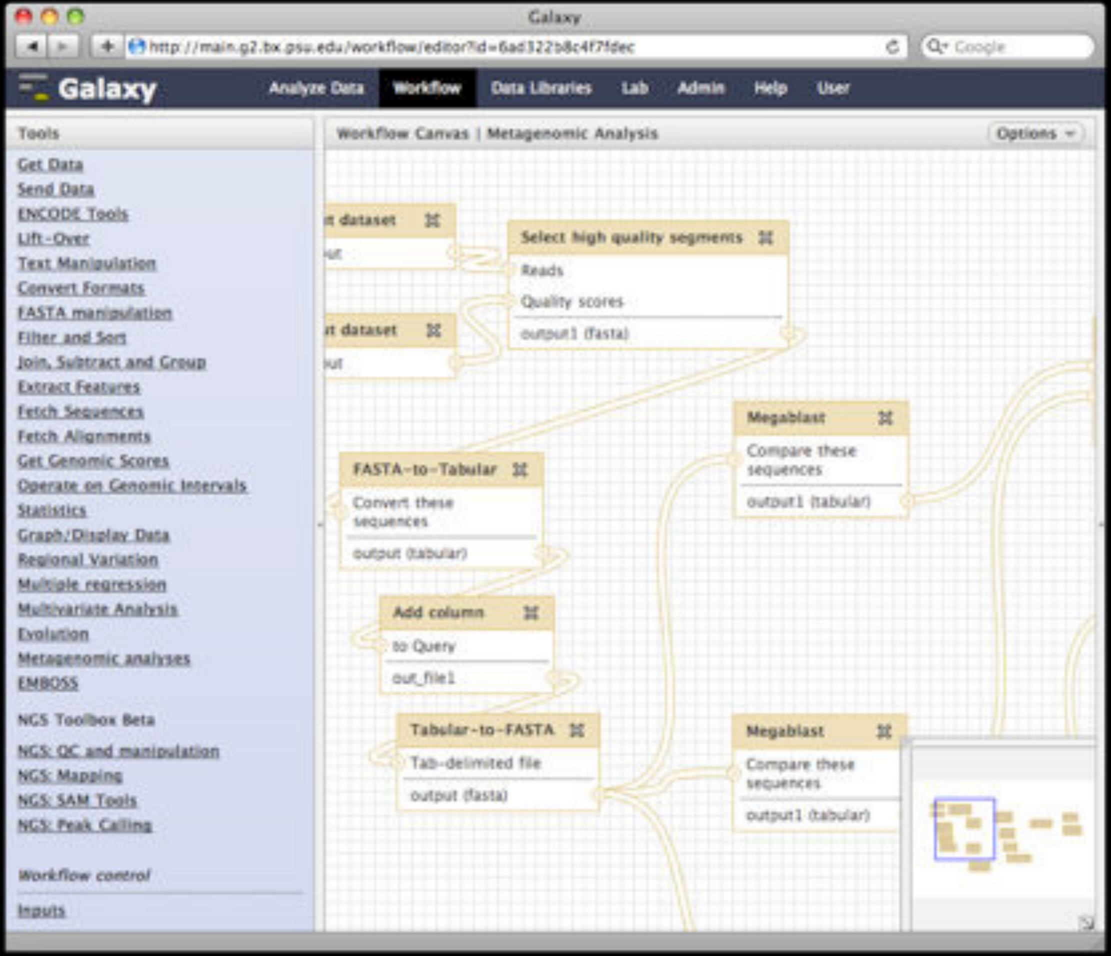
Annotation:

Find variants with coverage >= 30 and quality score >= 20.

| display at UCSC main | view in GeneTrack | display at Ensembl Current

1.Chrom	2.Start	3.End	4	5	6
chr10	6882036	6882037	A	A	107
chr10	14243075	14243076	G	G	96
chr10	14243079	14243080	C	C	106
chr10	14465082	14465083	T	K	173
chr10	14465083	14465084	G	K	144
chr10	14465084	14465085	T	T	117

Galaxy workflow system

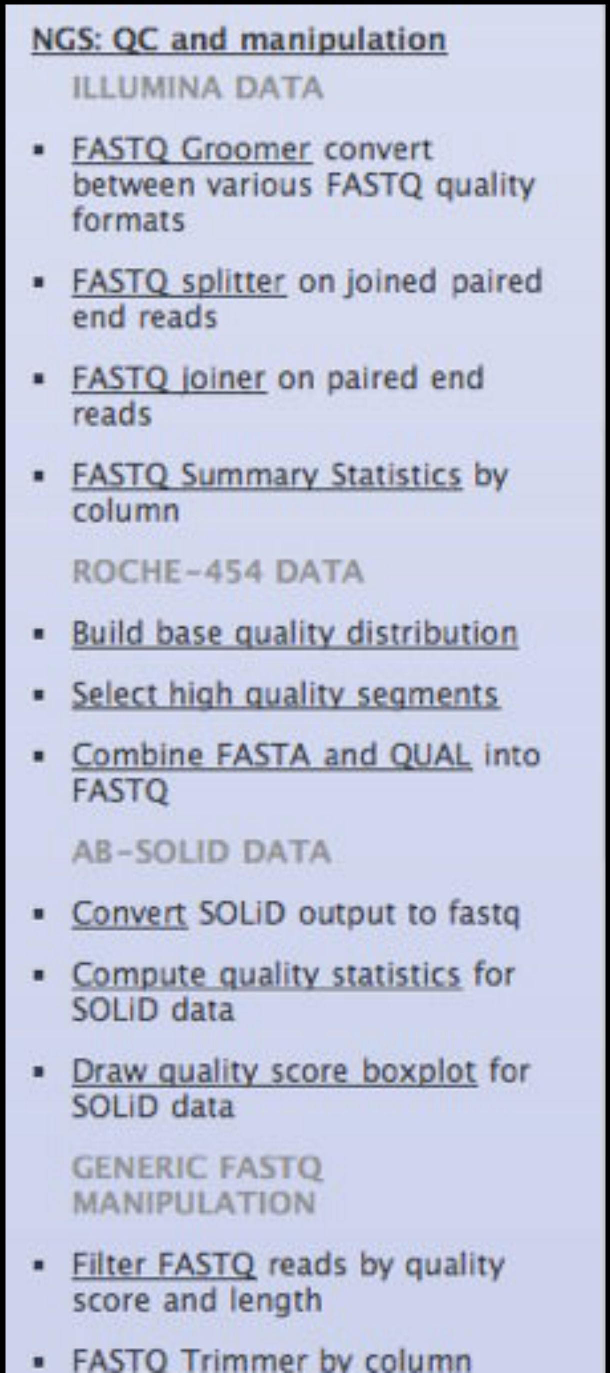


- Workflows can be constructed from scratch *or* extracted from existing analysis histories
- Facilitate reuse, as well as providing precise reproducibility of a complex analysis

Analyzing high throughput sequence data with Galaxy

- The Galaxy framework is generic; supporting a new type of analysis is as simple as integrating tools
- Galaxy is well suited to large-scale analysis
  - Allows tools to work with data in native, efficient formats
  - Integrates easily with cluster computing resources

(some) Galaxy tools for sequence data analysis



NGS: QC and manipulation

ILLUMINA DATA

- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

ROCHE-454 DATA

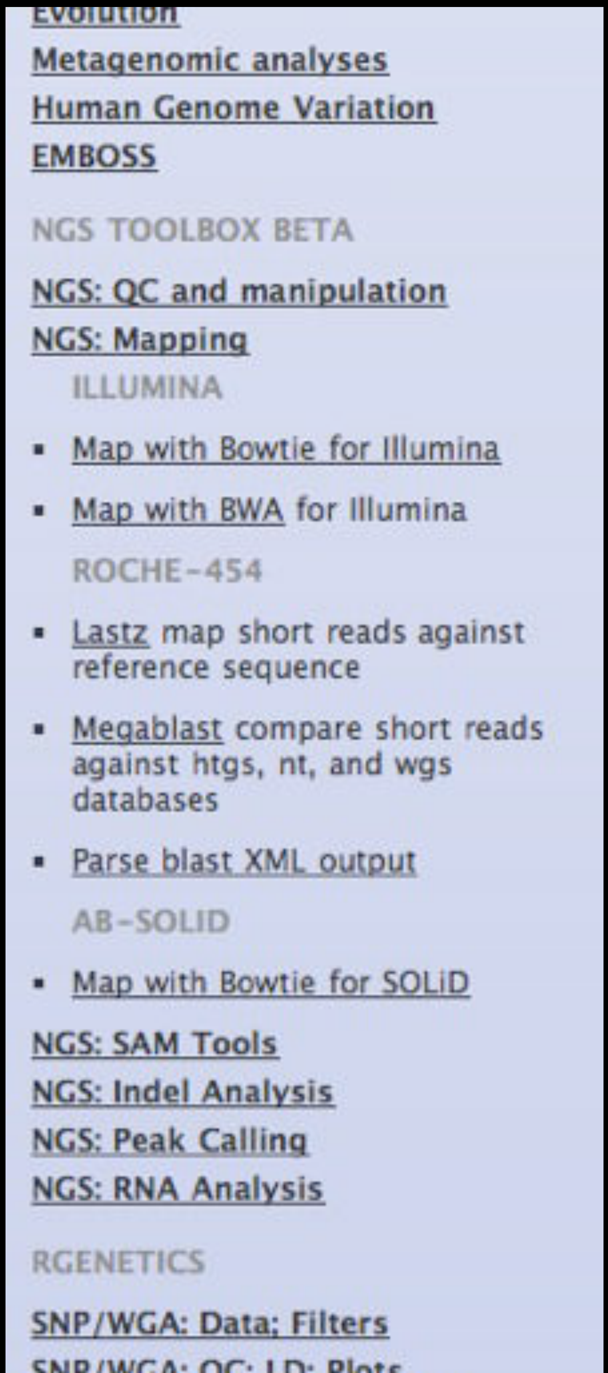
- Build base quality distribution
- Select high quality segments
- Combine FASTA and QUAL into FASTQ

AB-SOLID DATA

- Convert SOLiD output to fastq
- Compute quality statistics for SOLiD data
- Draw quality score boxplot for SOLiD data

GENERIC FASTQ MANIPULATION

- Filter FASTQ reads by quality score and length
- FASTQ Trimmer by column



Evolution

Metagenomic analyses

Human Genome Variation

EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation

NGS: Mapping

ILLUMINA

- Map with Bowtie for Illumina
- Map with BWA for Illumina

ROCHE-454

- Lastz map short reads against reference sequence
- Megablast compare short reads against htgs, nt, and wgs databases
- Parse blast XML output

AB-SOLID

- Map with Bowtie for SOLiD

NGS: SAM Tools

NGS: Indel Analysis

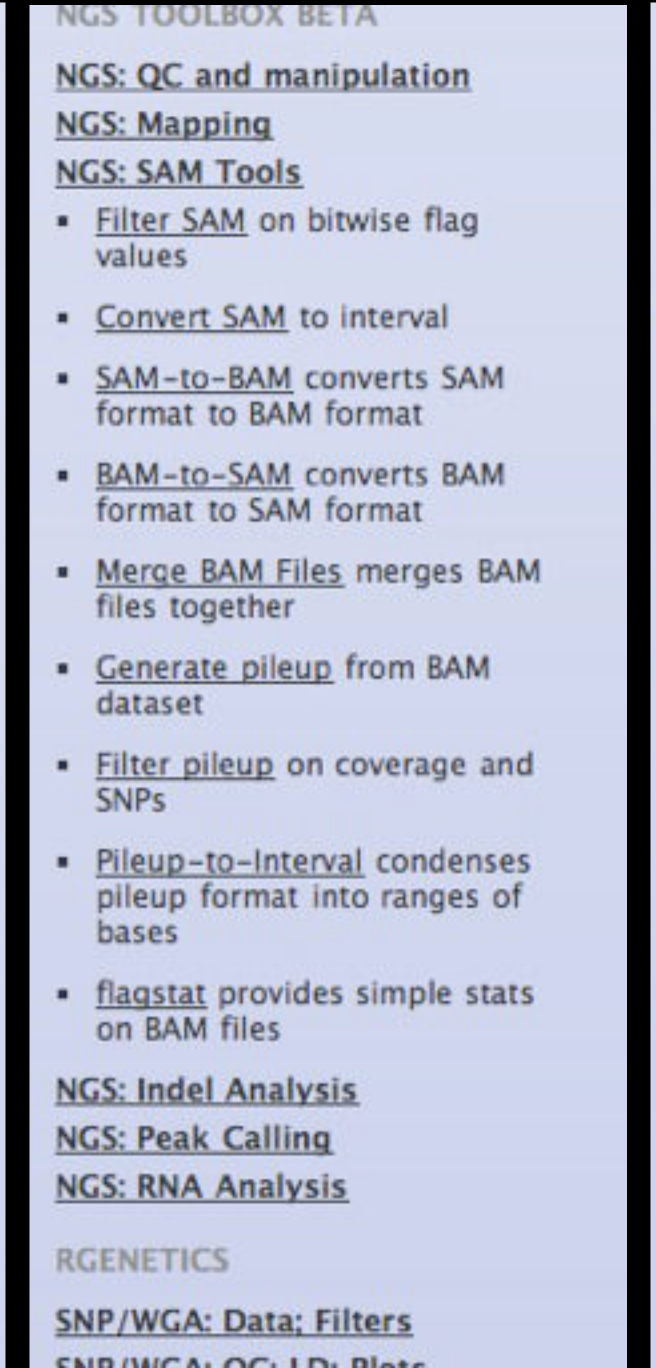
NGS: Peak Calling

NGS: RNA Analysis

RGENETICS

SNP/WGA: Data; Filters

SNP/WGA: QC; LD; Plots



NGS TOOLBOX BETA

NGS: QC and manipulation

NGS: Mapping

NGS: SAM Tools

- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format
- Merge BAM Files merges BAM files together
- Generate pileup from BAM dataset
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases
- flagstat provides simple stats on BAM files

NGS: Indel Analysis

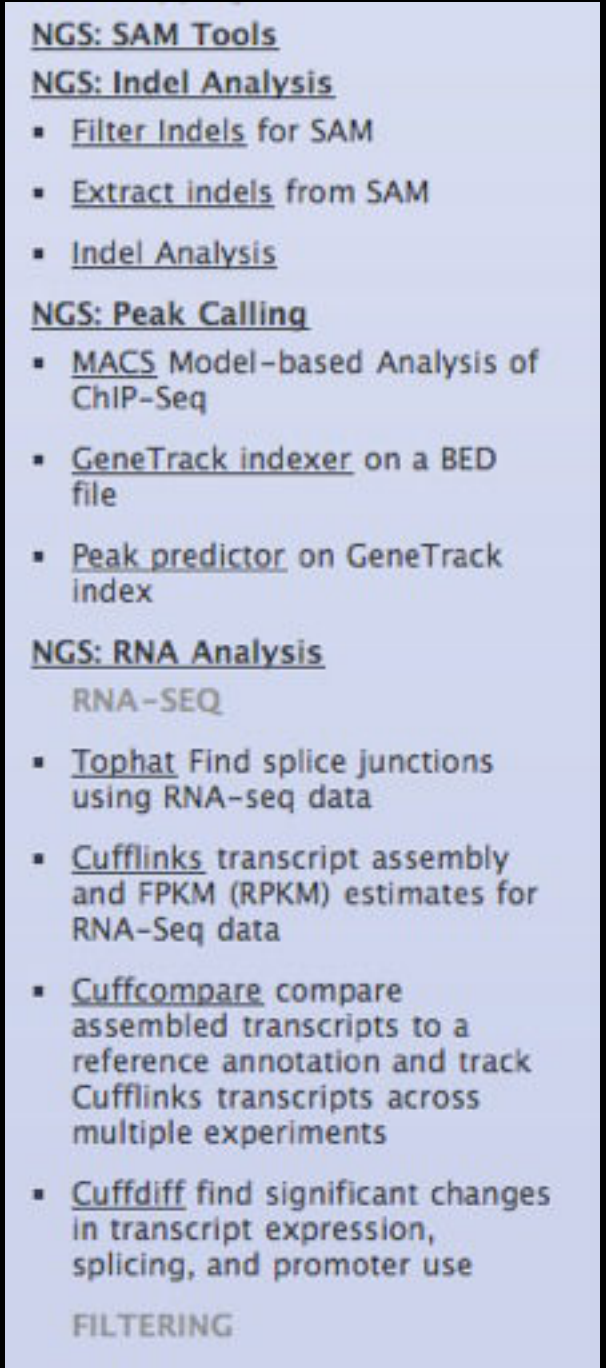
NGS: Peak Calling

NGS: RNA Analysis

RGENETICS

SNP/WGA: Data; Filters

SNP/WGA: QC; LD; Plots



NGS: SAM Tools

NGS: Indel Analysis

- Filter Indels for SAM
- Extract indels from SAM
- Indel Analysis

NGS: Peak Calling

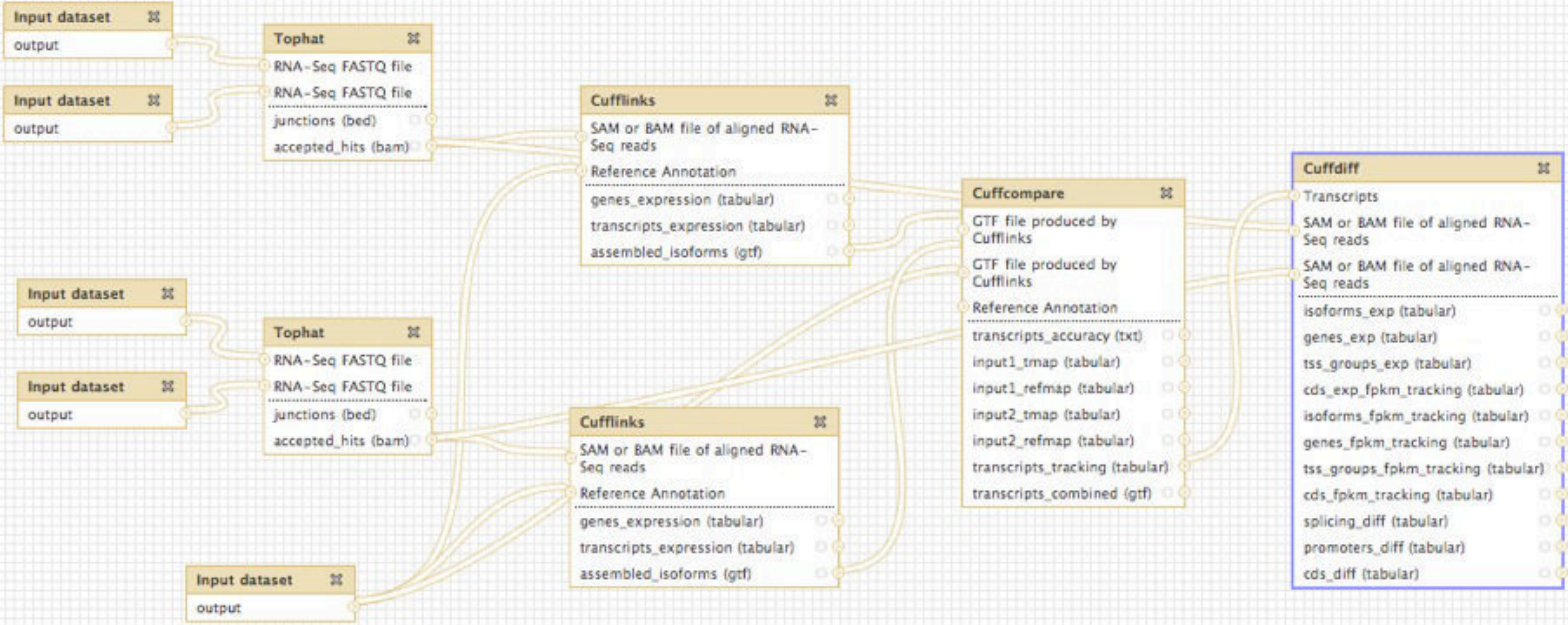
- MACS Model-based Analysis of ChIP-Seq
- GeneTrack indexer on a BED file
- Peak predictor on GeneTrack index

NGS: RNA Analysis

RNA-SEQ

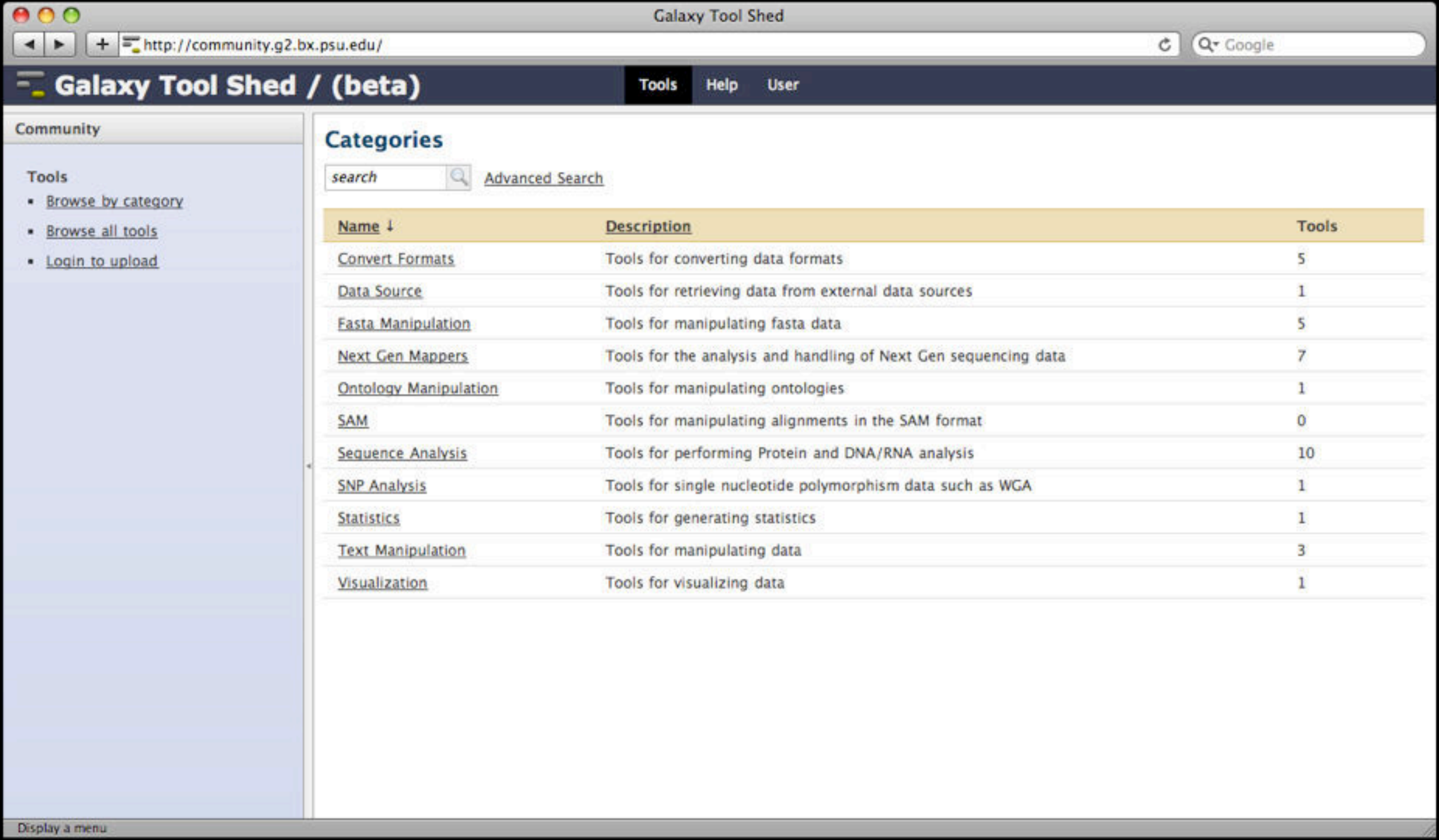
- Tophat Find splice junctions using RNA-seq data
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffdiff find significant changes in transcript expression, splicing, and promoter use

FILTERING



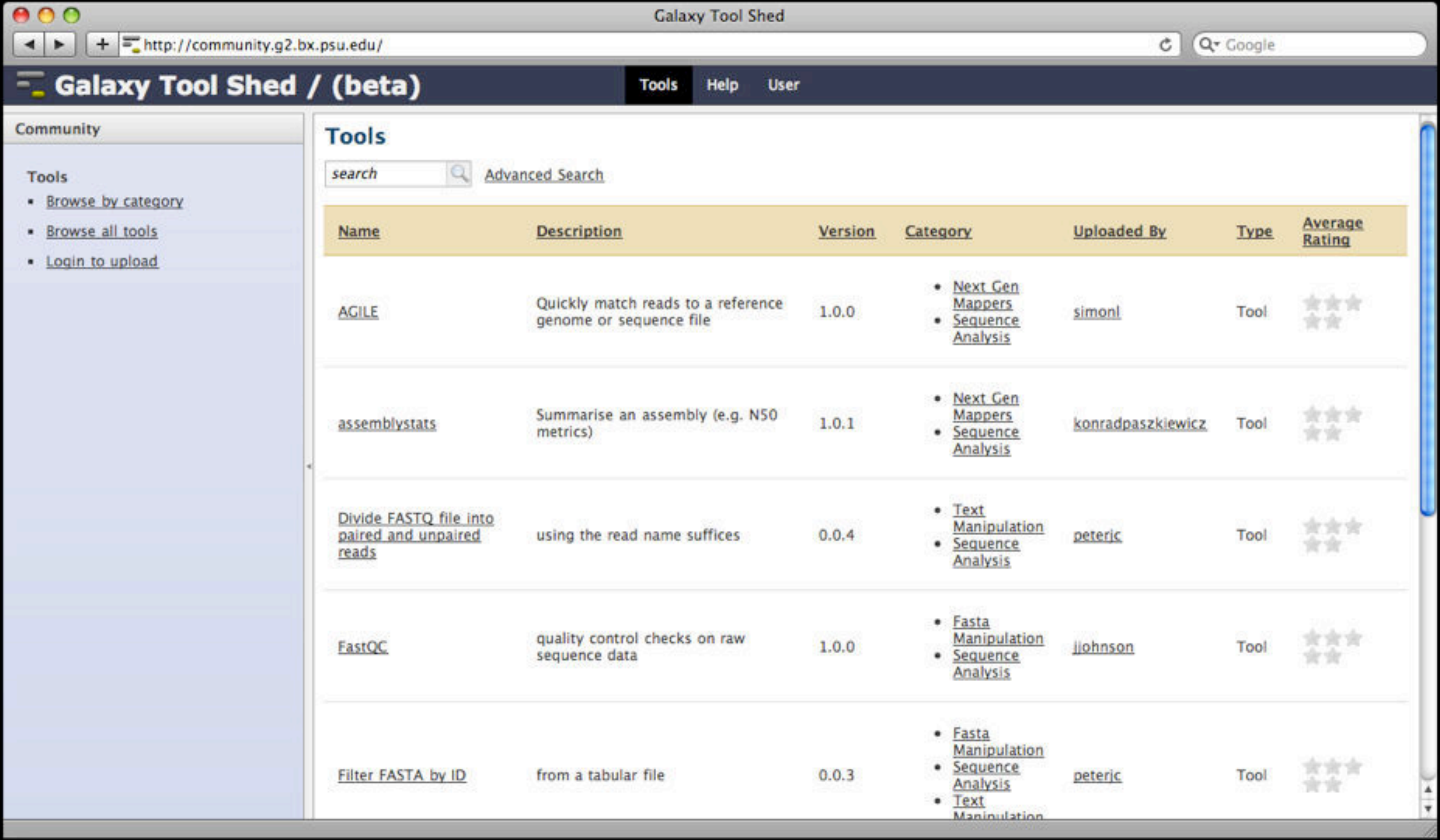
Example: Workflow for differential expression analysis of RNA-seq using Tophat/Cufflinks tools

Community of tool developers



Categories

Name ↓	Description	Tools
Convert Formats	Tools for converting data formats	5
Data Source	Tools for retrieving data from external data sources	1
Fasta Manipulation	Tools for manipulating fasta data	5
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	7
Ontology Manipulation	Tools for manipulating ontologies	1
SAM	Tools for manipulating alignments in the SAM format	0
Sequence Analysis	Tools for performing Protein and DNA/RNA analysis	10
SNP Analysis	Tools for single nucleotide polymorphism data such as WGA	1
Statistics	Tools for generating statistics	1
Text Manipulation	Tools for manipulating data	3
Visualization	Tools for visualizing data	1



Galaxy Tool Shed

http://community.g2.bx.psu.edu/

Galaxy Tool Shed / (beta)

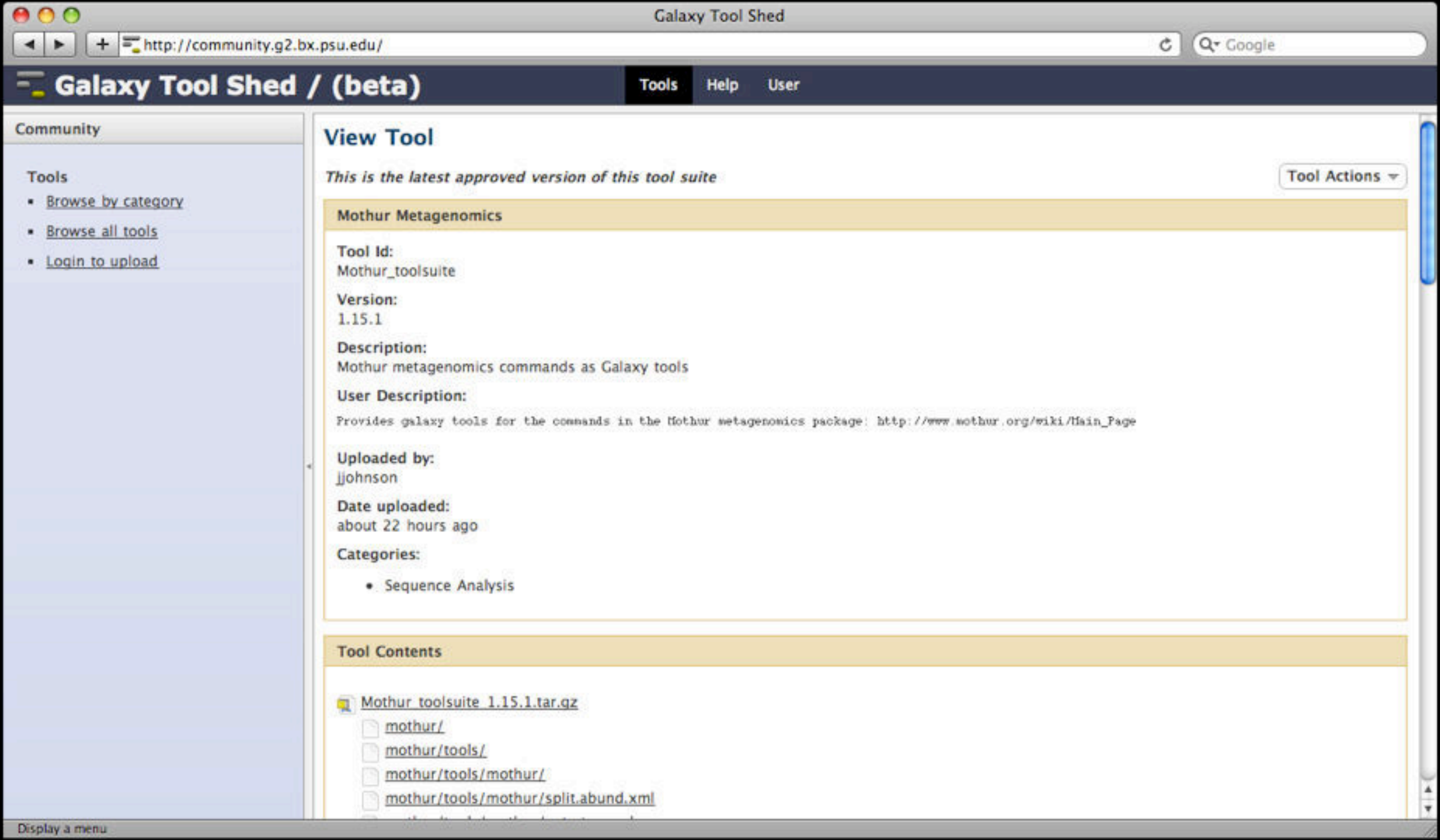
Community

Tools

- Browse by category
- Browse all tools
- Login to upload

Tools

Name	Description	Version	Category	Uploaded By	Type	Average Rating
AGILE	Quickly match reads to a reference genome or sequence file	1.0.0	• Next Gen Mappers • Sequence Analysis	simonl	Tool	☆☆☆☆☆
assemblystats	Summarise an assembly (e.g. N50 metrics)	1.0.1	• Next Gen Mappers • Sequence Analysis	konradpaszkiewicz	Tool	☆☆☆☆☆
Divide FASTQ file into paired and unpaired reads	using the read name suffices	0.0.4	• Text Manipulation • Sequence Analysis	peterjc	Tool	☆☆☆☆☆
FastQC	quality control checks on raw sequence data	1.0.0	• Fasta Manipulation • Sequence Analysis	jjohnson	Tool	☆☆☆☆☆
Filter FASTA by ID	from a tabular file	0.0.3	• Fasta Manipulation • Sequence Analysis • Text Manipulation	peterjc	Tool	☆☆☆☆☆



View Tool

This is the latest approved version of this tool suite

Mothur Metagenomics

Tool Id:
Mothur_toolsuite

Version:
1.15.1

Description:
Mothur metagenomics commands as Galaxy tools

User Description:

```
Provides galaxy tools for the commands in the Mothur metagenomics package: http://www.mothur.org/wiki/Main_Page
```

Uploaded by:
jjohnson

Date uploaded:
about 22 hours ago

Categories:

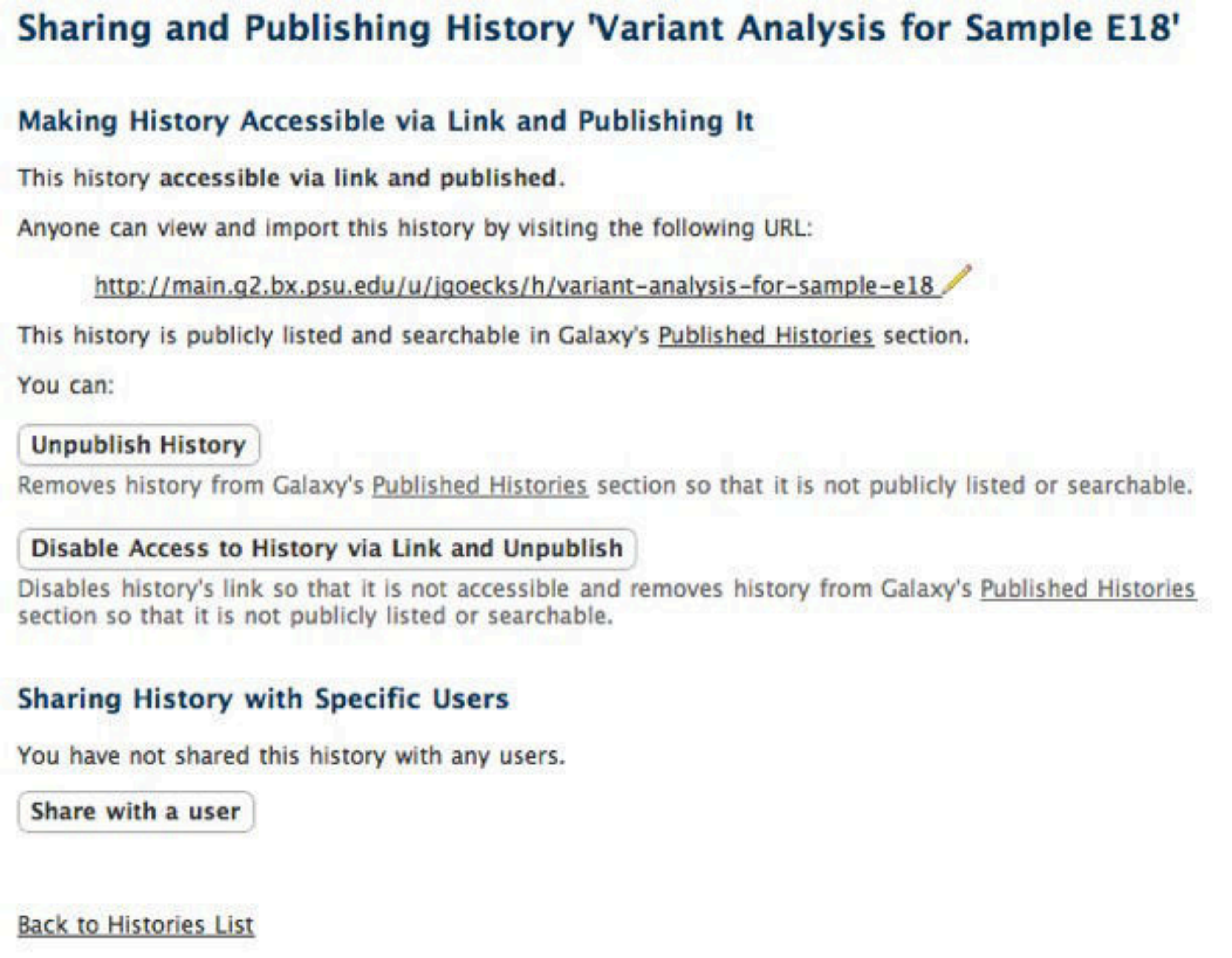
- Sequence Analysis

Tool Contents

Mothur_toolsuite_1.15.1.tar.gz
mothur/
mothur/tools/
mothur/tools/mothur/
mothur/tools/mothur/split.abund.xml

Data management

Everything can be shared and published



Sharing and Publishing History 'Variant Analysis for Sample E18'

Making History Accessible via Link and Publishing It

This history **accessible via link and published.**

Anyone can view and import this history by visiting the following URL:

http://main.g2.bx.psu.edu/u/jgoecks/h/variant-analysis-for-sample-e18

This history is publicly listed and searchable in Galaxy's Published Histories section.

You can:

Unpublish History

Removes history from Galaxy's Published Histories section so that it is not publicly listed or searchable.

Disable Access to History via Link and Unpublish

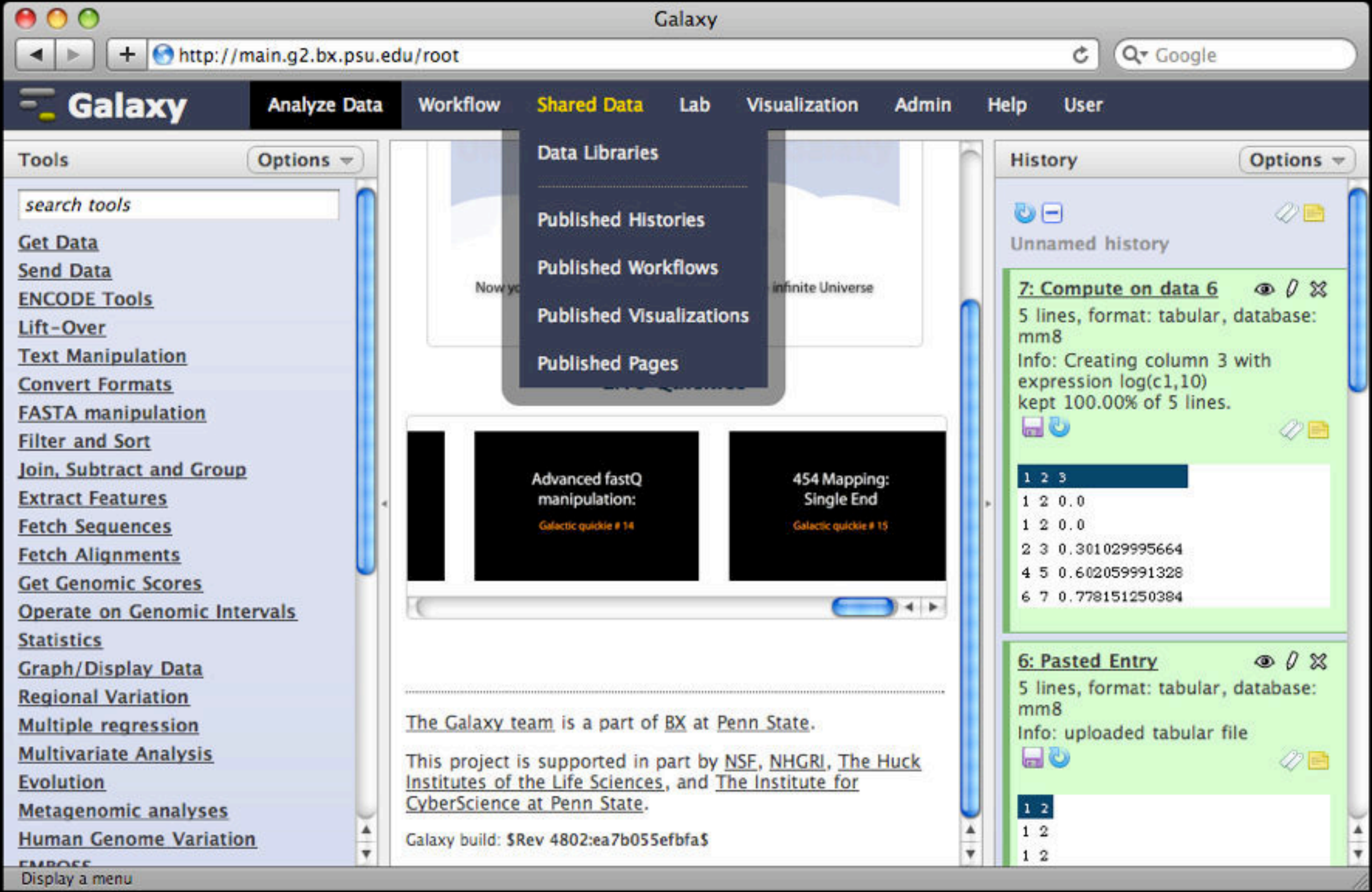
Disables history's link so that it is not accessible and removes history from Galaxy's Published Histories section so that it is not publicly listed or searchable.

Sharing History with Specific Users

You have not shared this history with any users.

Share with a user

Back to Histories List



7: Compute on data 6

5 lines, format: tabular, database: mm8

Info: Creating column 3 with expression log(c1,10)
kept 100.00% of 5 lines.

```
1 2 3
1 2 0.0
1 2 0.0
2 3 0.301029995664
4 5 0.602059991328
6 7 0.778151250384
```

6: Pasted Entry

5 lines, format: tabular, database: mm8

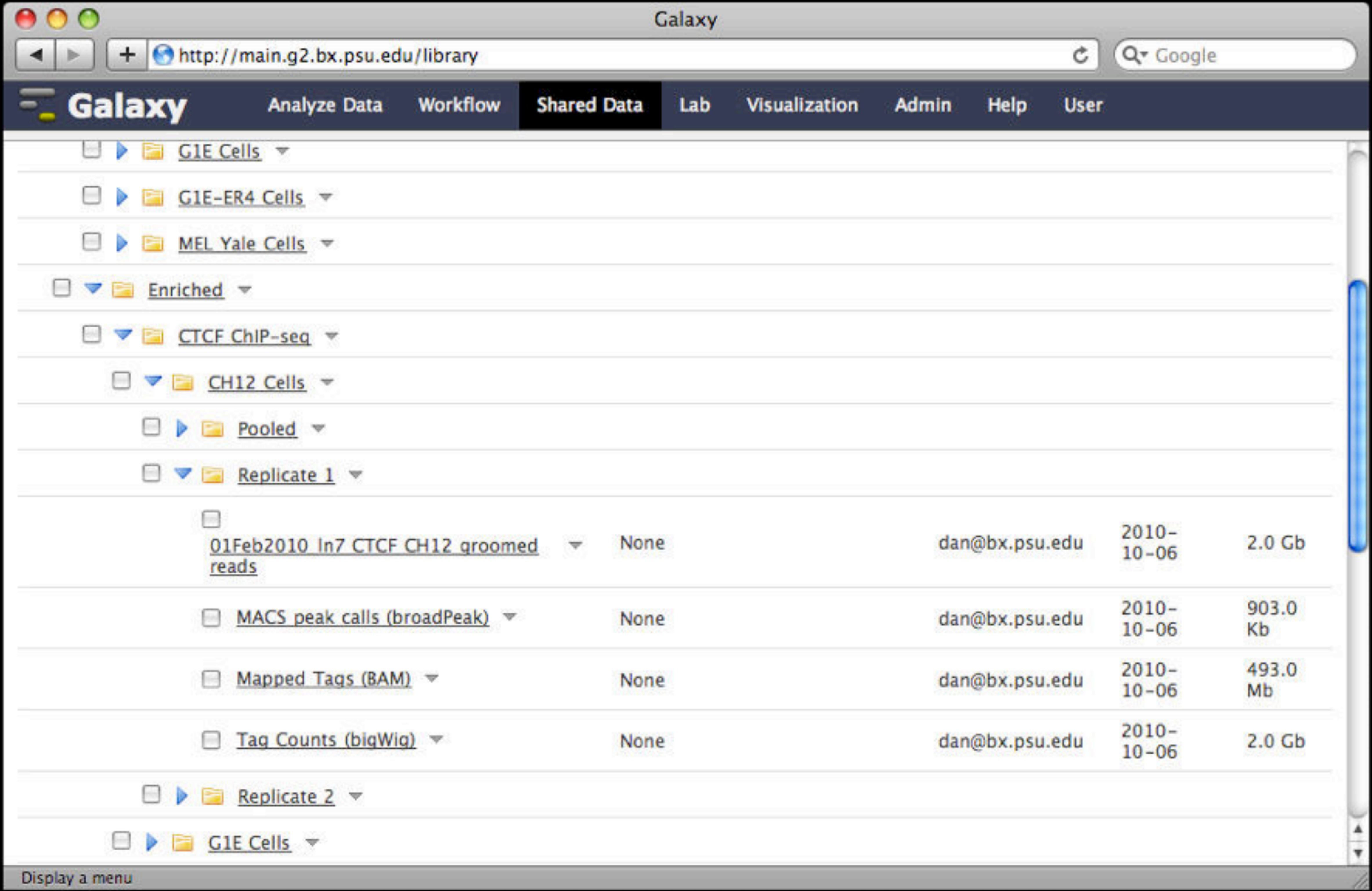
Info: uploaded tabular file

```
1 2
1 2
1 2
```

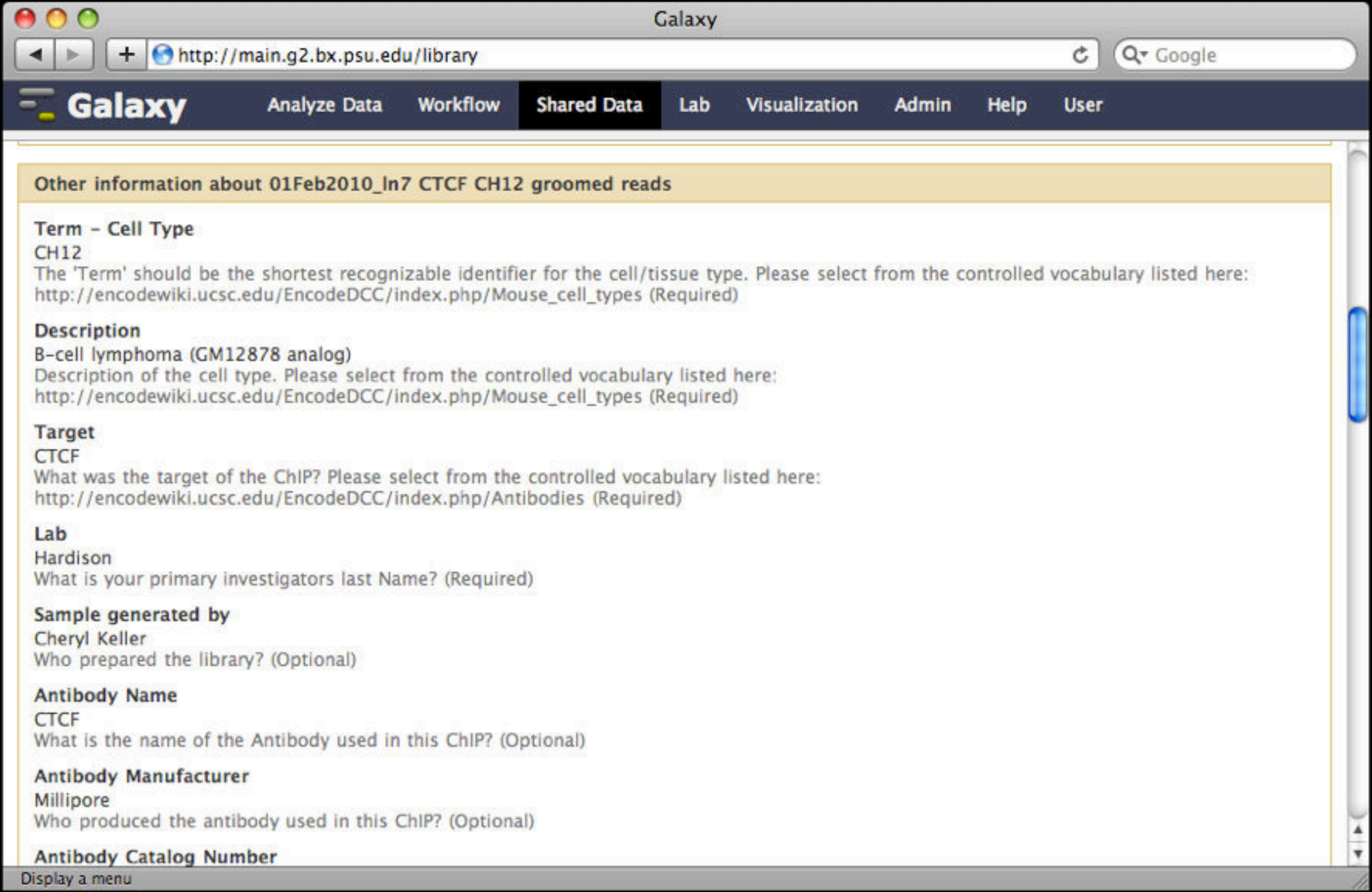
The Galaxy team is a part of BX at Penn State.

This project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, and The Institute for CyberScience at Penn State.

Galaxy build: $Rev 4802:ea7b055efbfa$



□ ▶ G1E Cells				
□ ▶ G1E-ER4 Cells				
□ ▶ MEL Yale Cells				
□ ▼ Enriched				
□ ▼ CTCF ChIP-seq				
□ ▼ CH12 Cells				
□ ▶ Pooled				
□ ▼ Replicate 1				
□ 01Feb2010 ln7 CTCF CH12 groomed reads	None	dan@bx.psu.edu	2010-10-06	2.0 Gb
□ MACS peak calls (broadPeak)	None	dan@bx.psu.edu	2010-10-06	903.0 Kb
□ Mapped Tags (BAM)	None	dan@bx.psu.edu	2010-10-06	493.0 Mb
□ Tag Counts (bigWig)	None	dan@bx.psu.edu	2010-10-06	2.0 Gb
□ ▶ Replicate 2				
□ ▶ G1E Cells				



Other information about 01Feb2010_ln7 CTCF CH12 groomed reads

Term - Cell Type
CH12
The 'Term' should be the shortest recognizable identifier for the cell/tissue type. Please select from the controlled vocabulary listed here: http://encodewiki.ucsc.edu/EncodeDCC/index.php/Mouse_cell_types (Required)

Description
B-cell lymphoma (GM12878 analog)
Description of the cell type. Please select from the controlled vocabulary listed here: http://encodewiki.ucsc.edu/EncodeDCC/index.php/Mouse_cell_types (Required)

Target
CTCF
What was the target of the ChIP? Please select from the controlled vocabulary listed here: http://encodewiki.ucsc.edu/EncodeDCC/index.php/Antibodies (Required)

Lab
Hardison
What is your primary investigators last Name? (Required)

Sample generated by
Cheryl Keller
Who prepared the library? (Optional)

Antibody Name
CTCF
What is the name of the Antibody used in this ChIP? (Optional)

Antibody Manufacturer
Millipore
Who produced the antibody used in this ChIP? (Optional)

Antibody Catalog Number

Making Galaxy your own

Building local Galaxy instances

- Galaxy is designed for local installation and customization
  - Just download and run, completely self-contained
  - Easily integrate new tools
  - Easy to deploy and manage on nearly any (unix) system
  - Run jobs on existing compute clusters

Scale up on your cluster

- Move intensive processing (tool execution) to other hosts
- Frees up the application server to serve requests and manage jobs
- Utilize existing resources
- Supports any scheduler that supports DRMAA (most of them)
- It's easy
- But, requires an **existing computational resource** on which to be deployed

CLUSTER RESOURCES™
TORQUE

GRID ENGINE

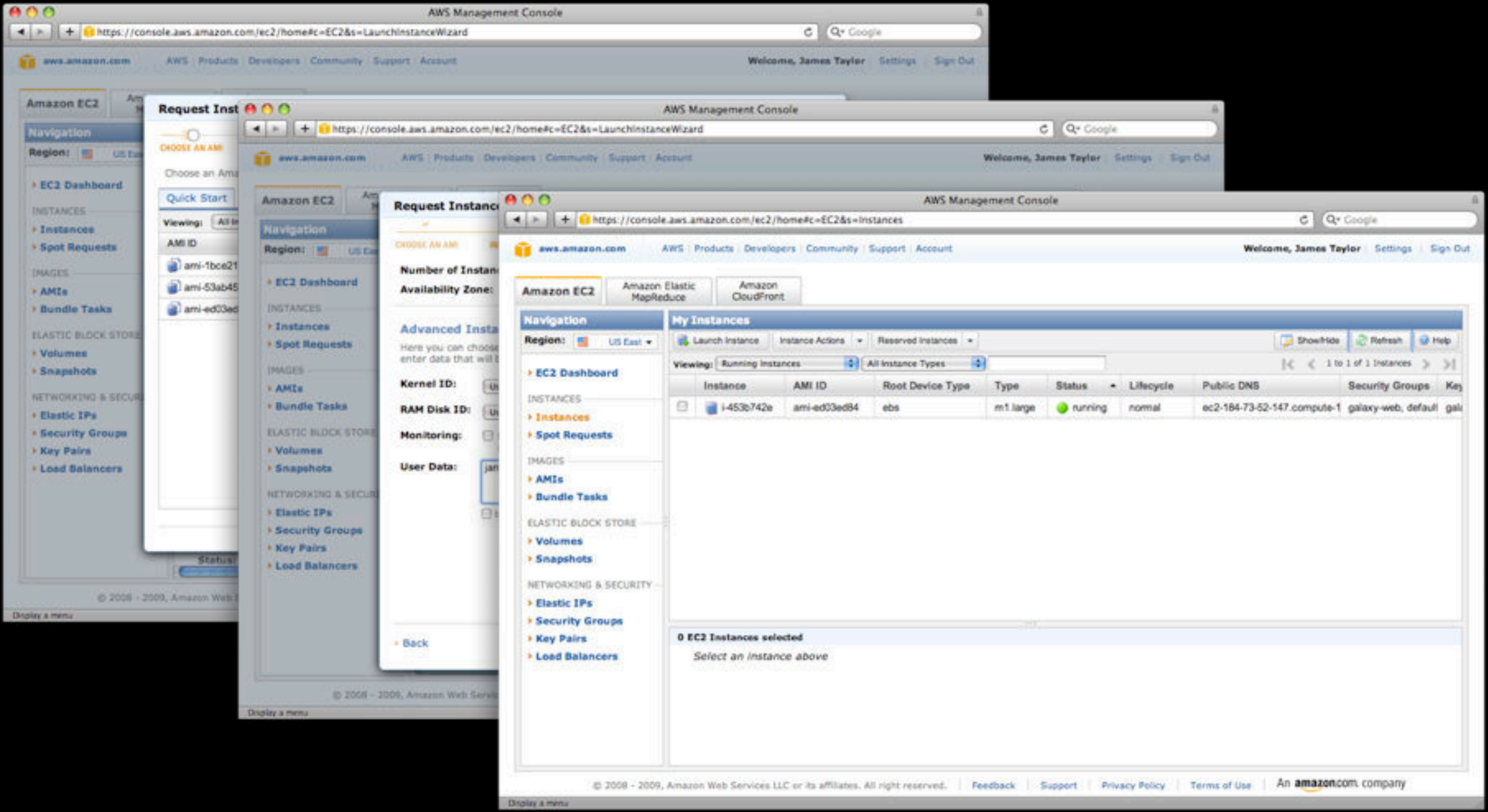
Platform Computing

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

Cloud computing / Infrastructure virtualization

- Computing using resources acquired on demand
- Virtual infrastructure allows for (potential) economies of scale, and (definite) improvements to management automation
- Cloud-style deployment provides a solution both for users without dedicated compute resources, and for simplifying deployment and management

Using Amazon EC2: Startup in 3 steps



https://console.aws.amazon.com/ec2/home#c=EC2&s=LaunchInstanceWizard

https://console.aws.amazon.com/ec2/home#c=EC2&s=Instances

Welcome, James Taylor

My Instances

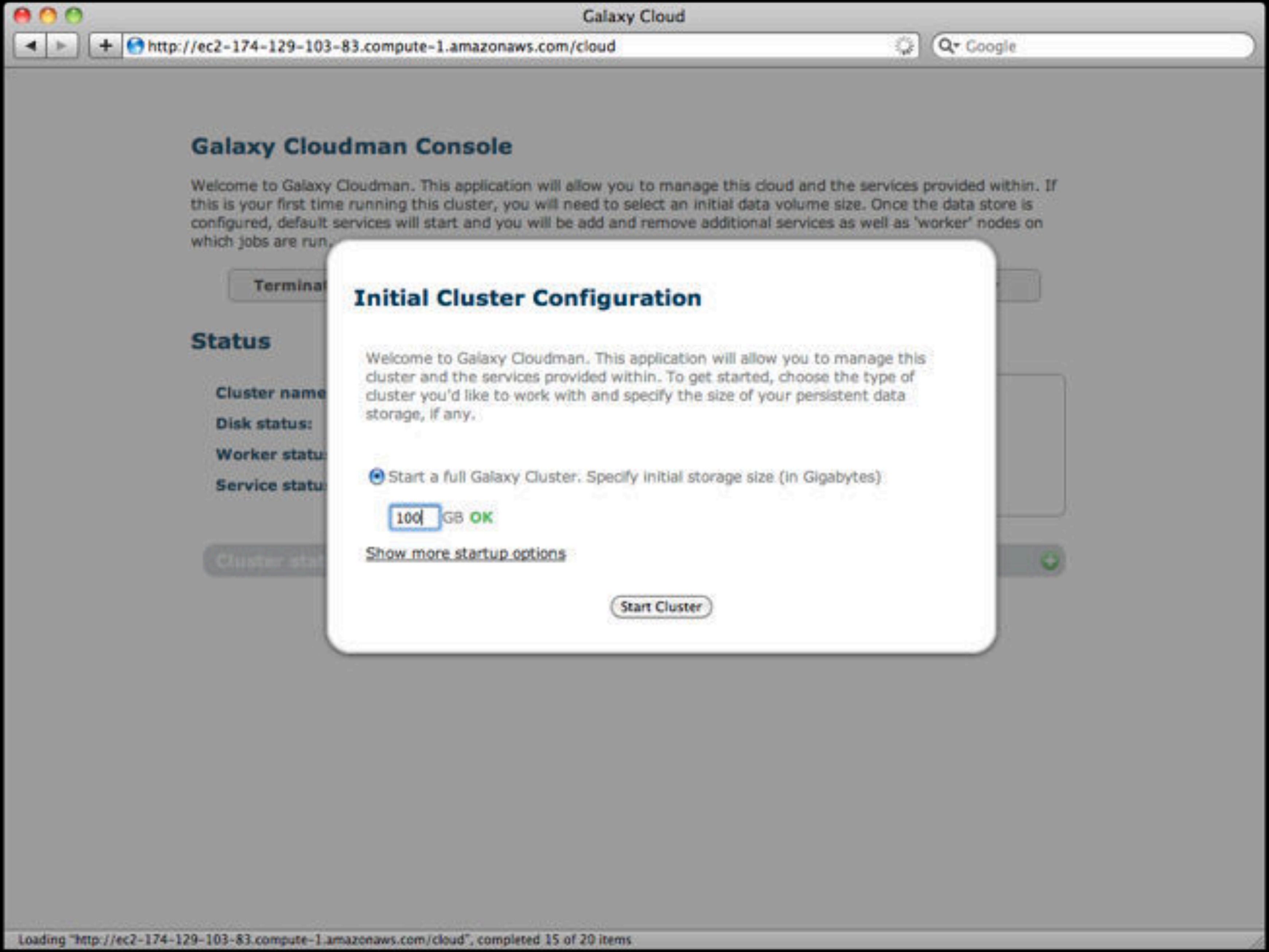
Viewing: Running Instances | All Instance Types

1 to 1 of 1 Instances

	Instance	AMI ID	Root Device Type	Type	Status	Lifecycle	Public DNS	Security Groups	Key
□	i-453b742e	ami-ed03ed84	ebs	m1.large	● running	normal	ec2-184-73-52-147.compute-1	galaxy-web, default	gala

0 EC2 Instances selected

Select an instance above



Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application will allow you to manage this cloud and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be add and remove additional services as well as 'worker' nodes on which jobs are run.

Status

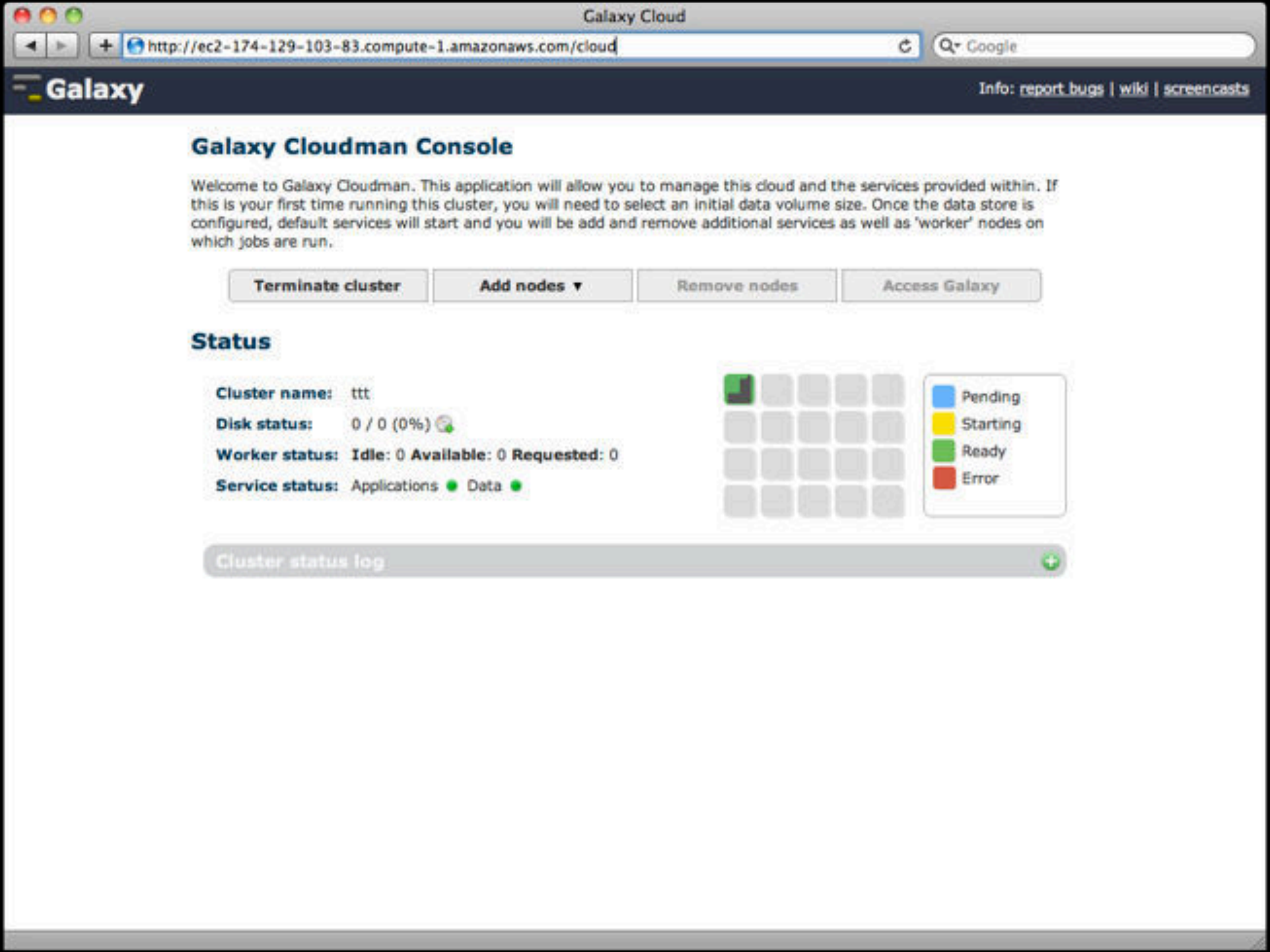
Initial Cluster Configuration

Welcome to Galaxy Cloudman. This application will allow you to manage this cluster and the services provided within. To get started, choose the type of cluster you'd like to work with and specify the size of your persistent data storage, if any.

● Start a full Galaxy Cluster. Specify initial storage size (in Gigabytes)

100 GB OK

Loading "http://ec2-174-129-103-83.compute-1.amazonaws.com/cloud", completed 15 of 20 items



Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application will allow you to manage this cloud and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be add and remove additional services as well as 'worker' nodes on which jobs are run.

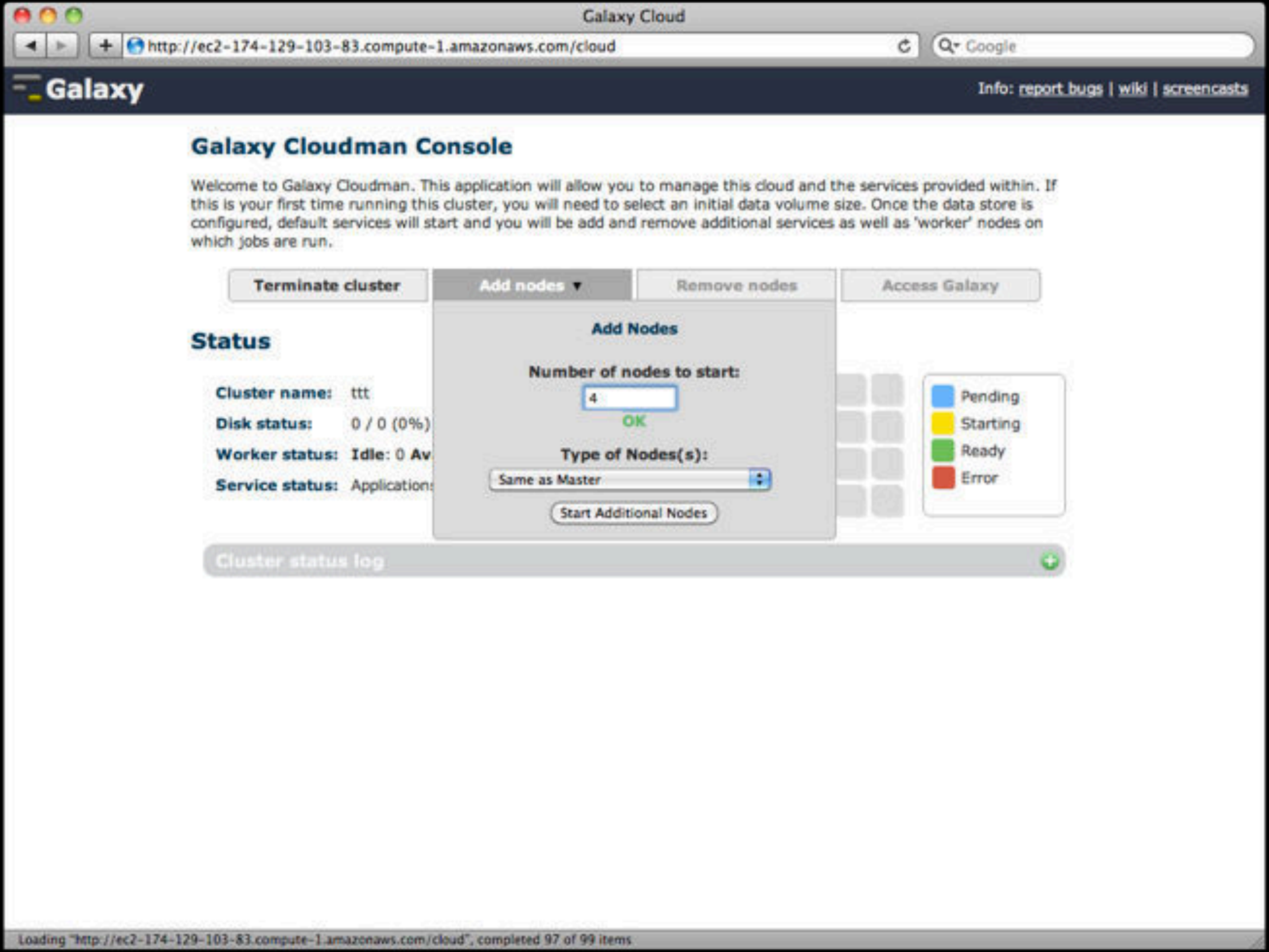
Status

Cluster name: ttt

Disk status: 0 / 0 (0%)

Worker status: **Idle**: 0 **Available**: 0 **Requested**: 0

Service status: Applications ● Data ●



Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application will allow you to manage this cloud and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be add and remove additional services as well as 'worker' nodes on which jobs are run.

Status

Cluster name: ttt

Disk status: 0 / 0 (0%)

Worker status: **Idle**: 0 **Av**

Service status: Application

Add Nodes

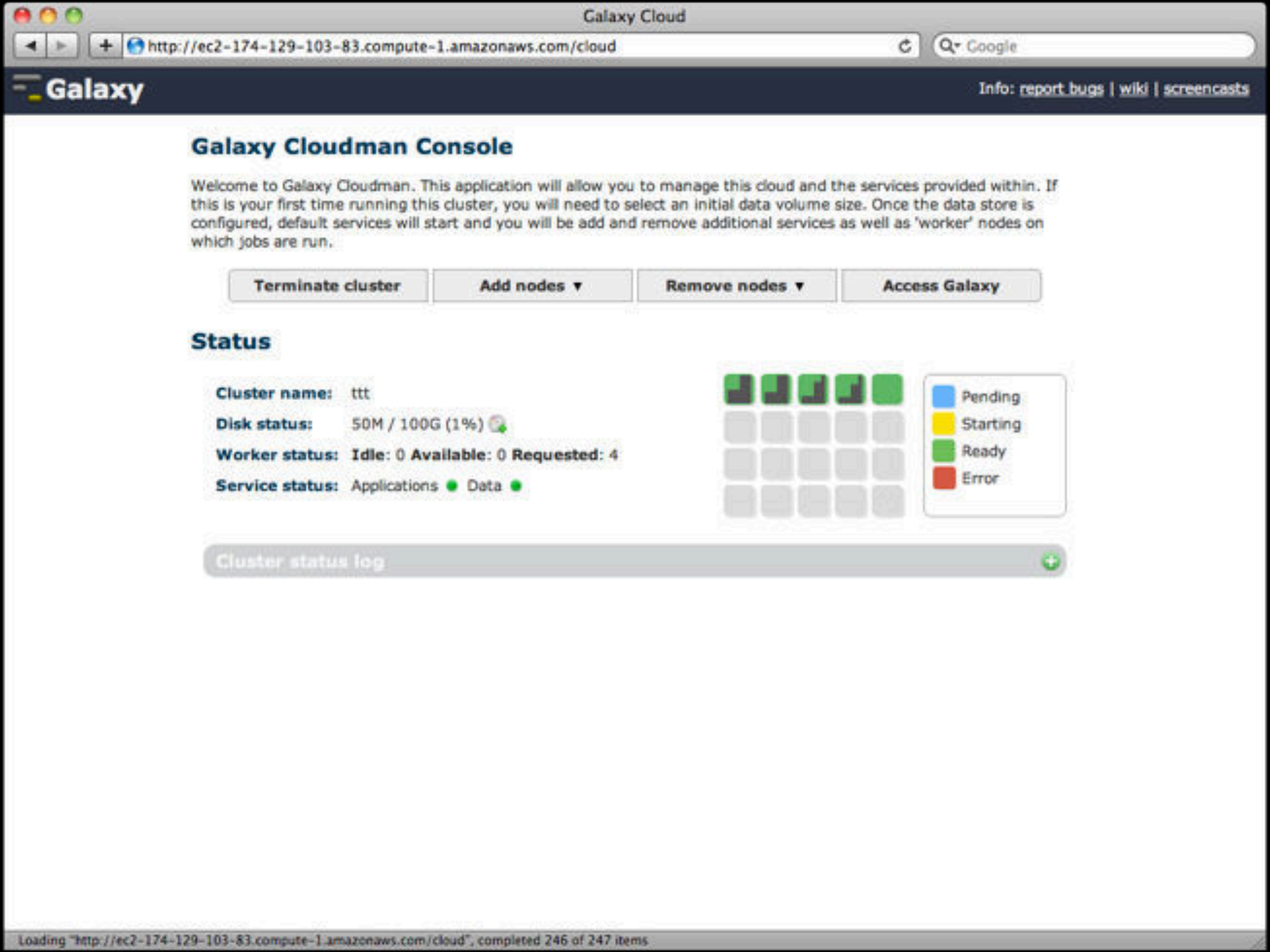
Number of nodes to start: 4

OK

Type of Nodes(s): Same as Master

🟦 Pending
🟨 Starting
🟩 Ready
🟥 Error

Loading "http://ec2-174-129-103-83.compute-1.amazonaws.com/cloud", completed 97 of 99 items



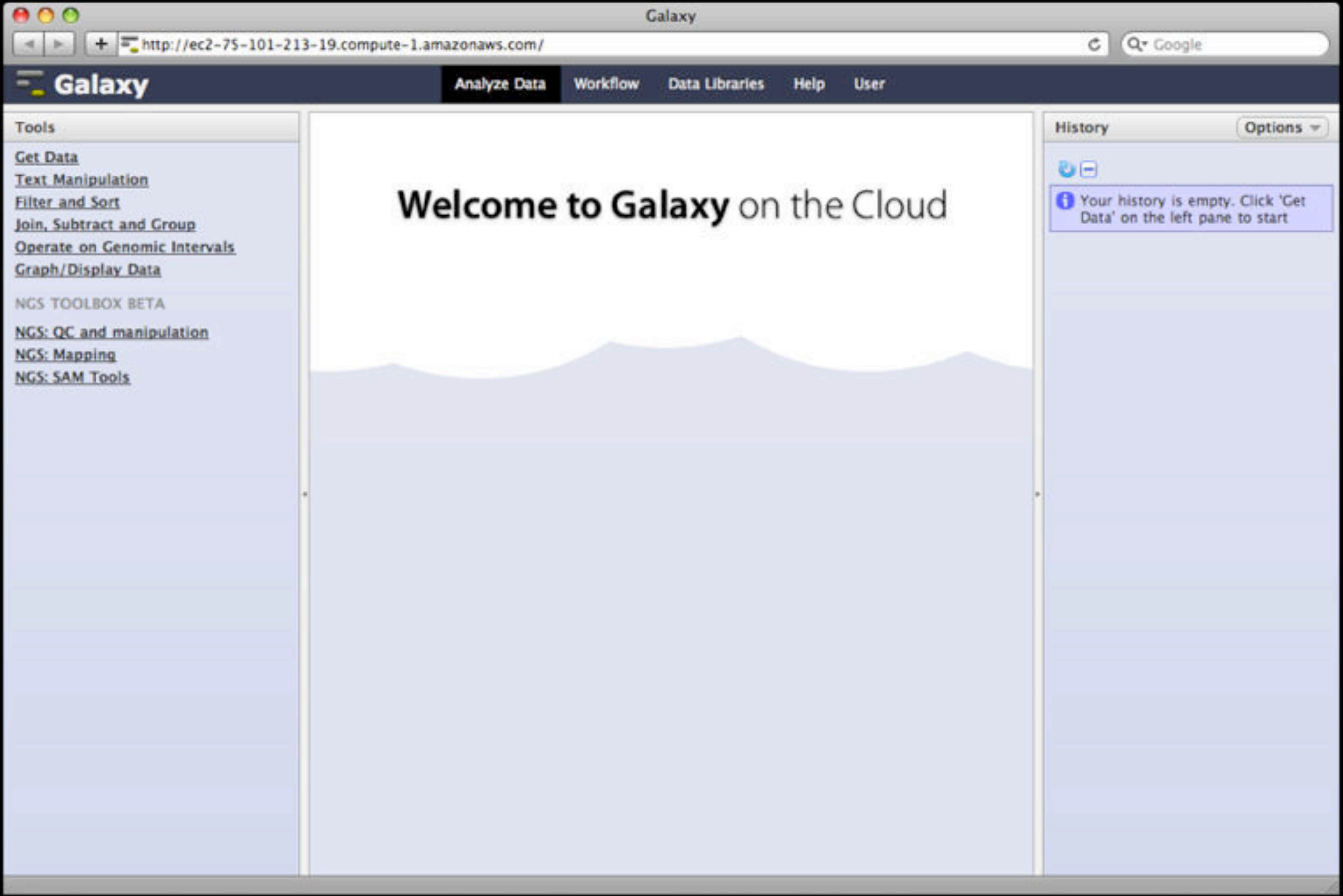
Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application will allow you to manage this cloud and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be add and remove additional services as well as 'worker' nodes on which jobs are run.

Status

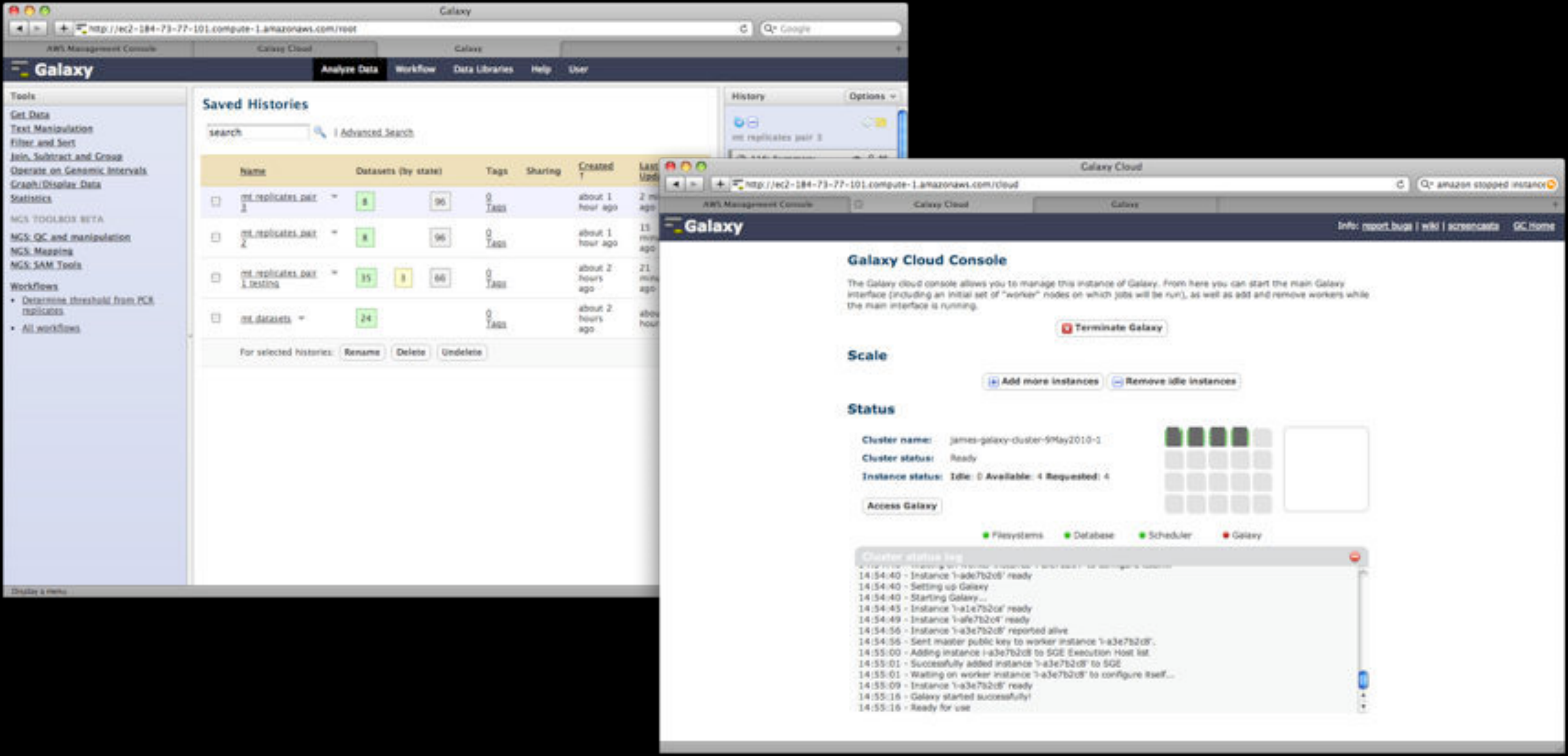
Cluster name:	ttt
Disk status:	50M / 100G (1%)
Worker status:	**Idle**: 0 **Available**: 0 **Requested**: 4
Service status:	Applications ● Data ●

Loading "http://ec2-174-129-103-83.compute-1.amazonaws.com/cloud", completed 246 of 247 items



Welcome to Galaxy on the Cloud

Your history is empty. Click 'Get Data' on the left pane to start



Saved Histories

search | Advanced Search

	Name	Datasets (by state)			Tags	Sharing	Created	Last Upd
□	mt replicates pair 3	8		96	0 Tags		about 1 hour ago	2 m ago
□	mt replicates pair 2	8		96	0 Tags		about 1 hour ago	13 min ago
□	mt replicates pair 1 testing	35	3	66	0 Tags		about 2 hours ago	21 min ago
□	mt datasets	24			0 Tags		about 2 hours ago	abo hour

For selected histories: Rename Delete Undelete

Galaxy Cloud
http://ec2-184-73-77-101.compute-1.amazonaws.com/cloud
amazon stopped instance
AWS Management Console
Galaxy Cloud
Galaxy
Info: report bugs | wiki | screencasts
GC Home

Galaxy Cloud Console

The Galaxy cloud console allows you to manage this instance of Galaxy. From here you can start the main Galaxy interface (including an initial set of "worker" nodes on which jobs will be run), as well as add and remove workers while the main interface is running.

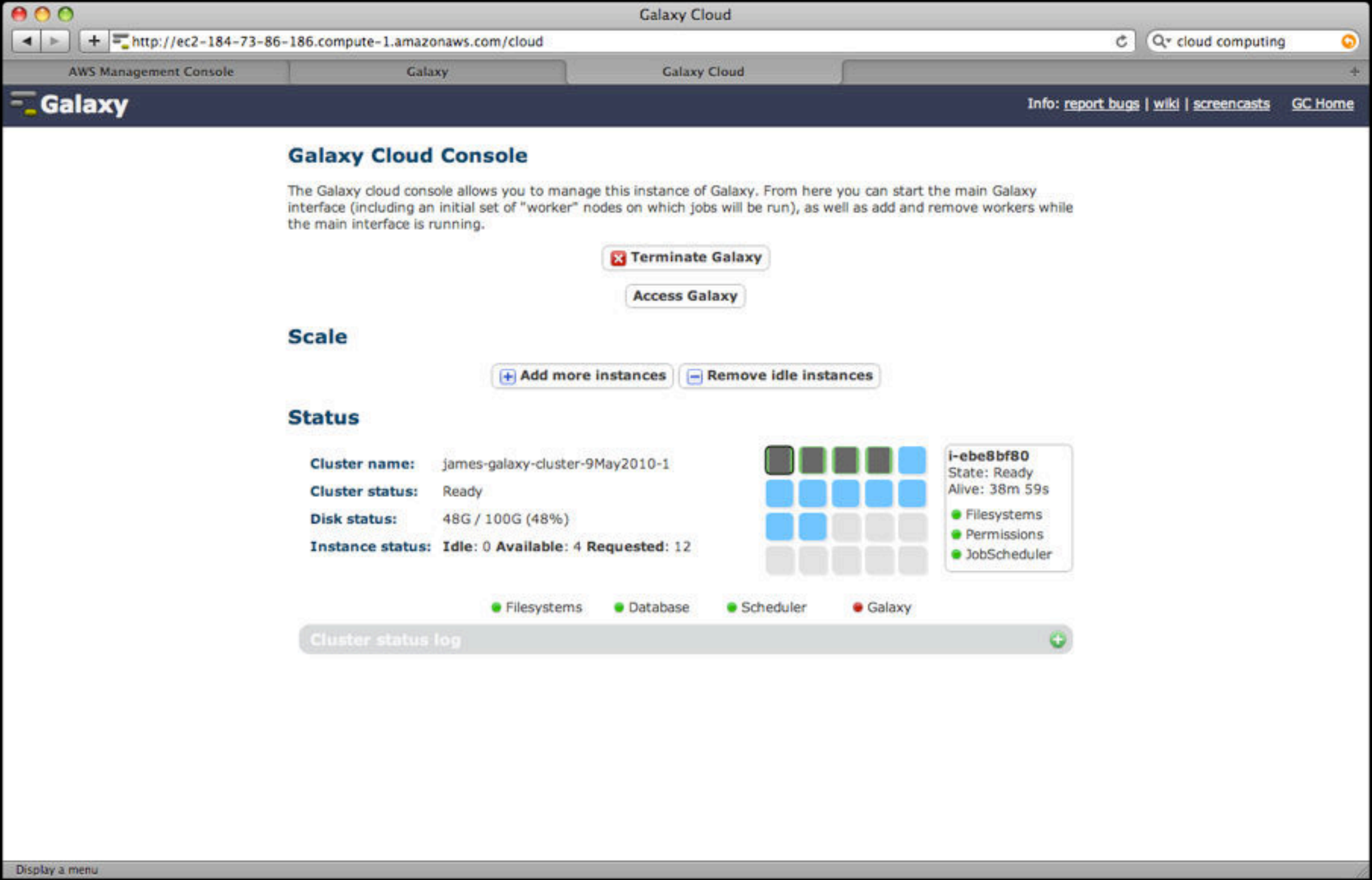
Scale

Status

Cluster name: james-galaxy-cluster-9May2010-1
Cluster status: Ready
Instance status: Idle: 0 Available: 4 Requested: 4

```
14:54:40 - Instance 'i-ade7b2c6' ready
14:54:40 - Setting up Galaxy
14:54:40 - Starting Galaxy...
14:54:45 - Instance 'i-a1e7b2ca' ready
14:54:49 - Instance 'i-afe7b2c4' ready
14:54:56 - Instance 'i-a3e7b2c8' reported alive
14:54:56 - Sent master public key to worker instance 'i-a3e7b2c8'.
14:55:00 - Adding instance i-a3e7b2c8 to SGE Execution Host list
14:55:01 - Successfully added instance 'i-a3e7b2c8' to SGE
14:55:01 - Waiting on worker instance 'i-a3e7b2c8' to configure itself...
14:55:09 - Instance 'i-a3e7b2c8' ready
14:55:16 - Galaxy started successfully!
14:55:16 - Ready for use
```

Can use like any other Galaxy instance, with additional compute nodes acquired and released (*automatically*) in response to usage


Galaxy Cloud
http://ec2-184-73-86-186.compute-1.amazonaws.com/cloud
cloud computing
AWS Management Console
Galaxy
Galaxy Cloud
Galaxy
Info: report bugs | wiki | screencasts
GC Home

Galaxy Cloud Console

The Galaxy cloud console allows you to manage this instance of Galaxy. From here you can start the main Galaxy interface (including an initial set of "worker" nodes on which jobs will be run), as well as add and remove workers while the main interface is running.

Scale

Status

Cluster name: james-galaxy-cluster-9May2010-1

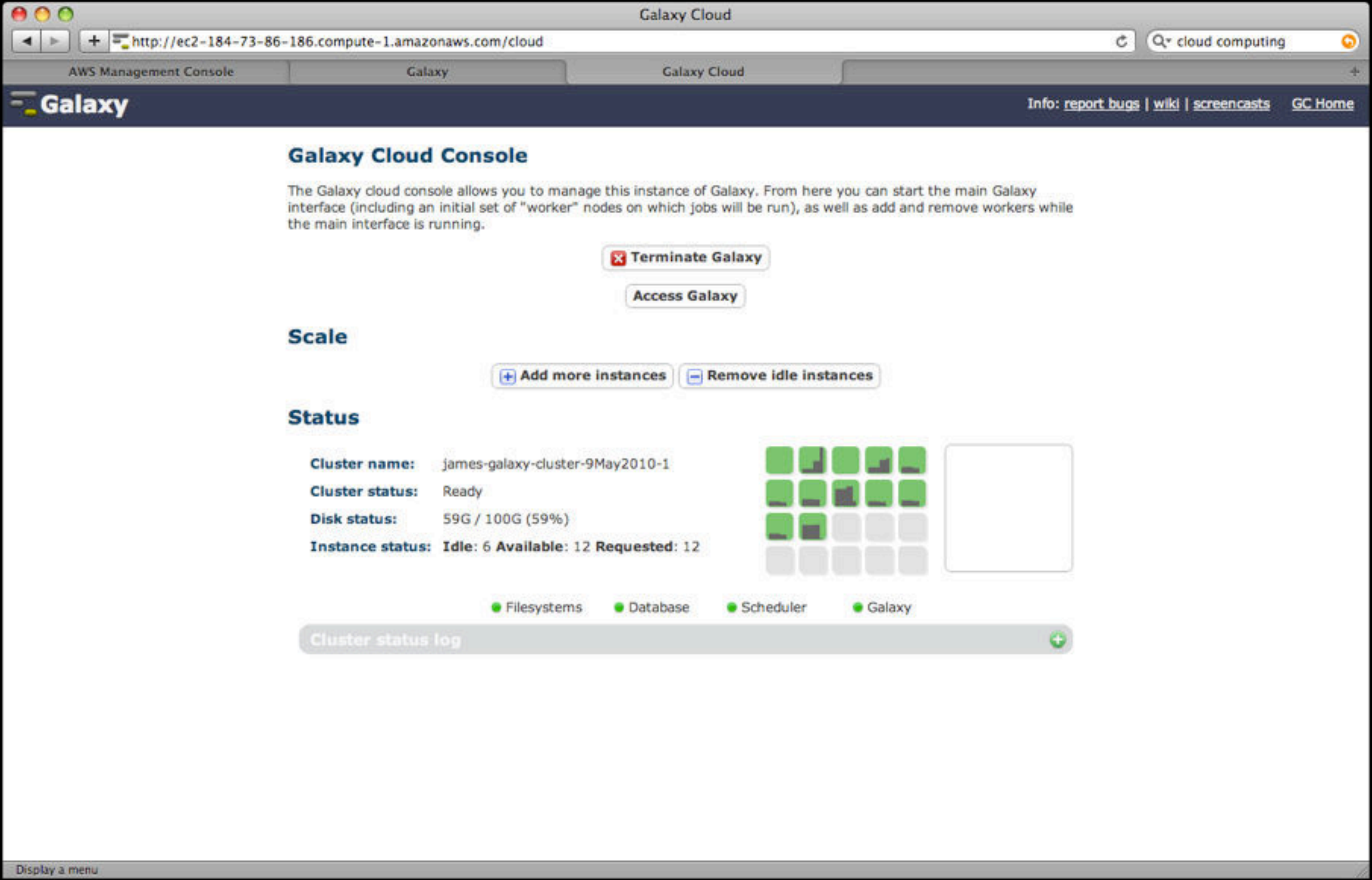
Cluster status: Ready

Disk status: 48G / 100G (48%)

Instance status: **Idle**: 0 **Available**: 4 **Requested**: 12

i-ebe8bf80
State: Ready
Alive: 38m 59s

- ● Filesystems
- ● Permissions
- ● JobScheduler



Galaxy Cloud
http://ec2-184-73-86-186.compute-1.amazonaws.com/cloud
cloud computing
AWS Management Console
Galaxy
Galaxy Cloud
Galaxy
Info: report bugs | wiki | screencasts
GC Home

Galaxy Cloud Console

The Galaxy cloud console allows you to manage this instance of Galaxy. From here you can start the main Galaxy interface (including an initial set of "worker" nodes on which jobs will be run), as well as add and remove workers while the main interface is running.

Scale

Status

Cluster name: james-galaxy-cluster-9May2010-1

Cluster status: Ready

Disk status: 59G / 100G (59%)

Instance status: **Idle**: 6 **Available**: 12 **Requested**: 12

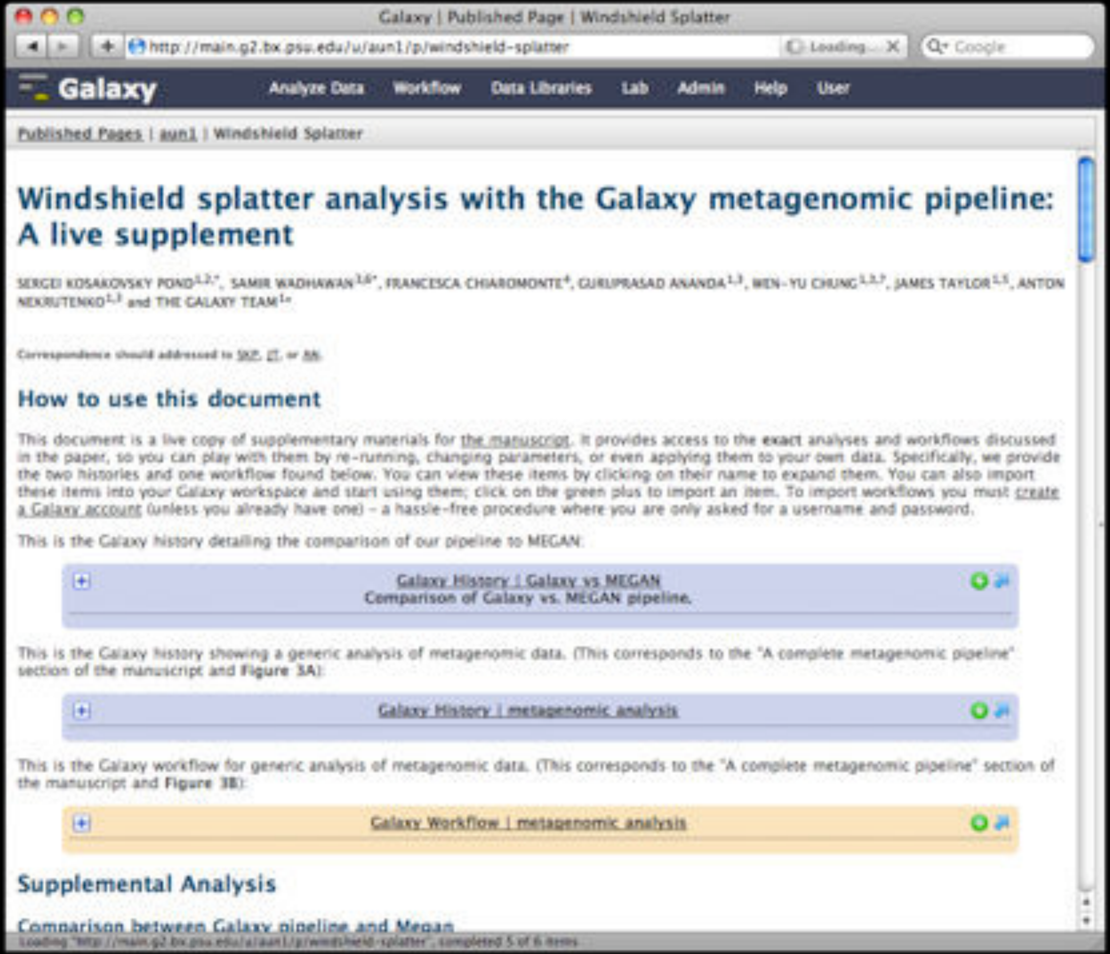
● Filesystems ● Database ● Scheduler ● Galaxy

Persistence

- Once analysis is complete, can scale down worker nodes or shutdown the entire analysis interface
- Data, configuration, *et cetera* is stored, and you can start the cluster back up to continue analysis at any time
  - Pay for just what you need

Publishing analysis

Sharing and publishing


Galaxy | Published Page | Windshield Splatter
http://main.g2.bx.psu.edu/u/aun1/p/windshield-splatter
Loading...
Google
Galaxy
Analyze Data
Workflow
Data Libraries
Lab
Admin
Help
User
Published Pages | aun1 | Windshield Splatter

Windshield splatter analysis with the Galaxy metagenomic pipeline: A live supplement

SERGEI KOSAKOVSKY POND[1,2,*], SAMIR WADHAWAN[1,6*], FRANCESCA CHIAROMONTE[4], GURUPRASAD ANANDA[1,3], WEN-YU CHUNG[1,3,7], JAMES TAYLOR[1,5], ANTON NEKRUTENKO[1,3] and THE GALAXY TEAM[1*]

Correspondence should addressed to SKP, JT, or AN.

How to use this document

This document is a live copy of supplementary materials for the manuscript. It provides access to the **exact** analyses and workflows discussed in the paper, so you can play with them by re-running, changing parameters, or even applying them to your own data. Specifically, we provide the two histories and one workflow found below. You can view these items by clicking on their name to expand them. You can also import these items into your Galaxy workspace and start using them; click on the green plus to import an item. To import workflows you must create a Galaxy account (unless you already have one) - a hassle-free procedure where you are only asked for a username and password.

This is the Galaxy history detailing the comparison of our pipeline to MEGAN:

Galaxy History | Galaxy vs MEGAN
Comparison of Galaxy vs. MEGAN pipeline.

This is the Galaxy history showing a generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3A**):

Galaxy History | metagenomic analysis

This is the Galaxy workflow for generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3B**):

Galaxy Workflow | metagenomic analysis

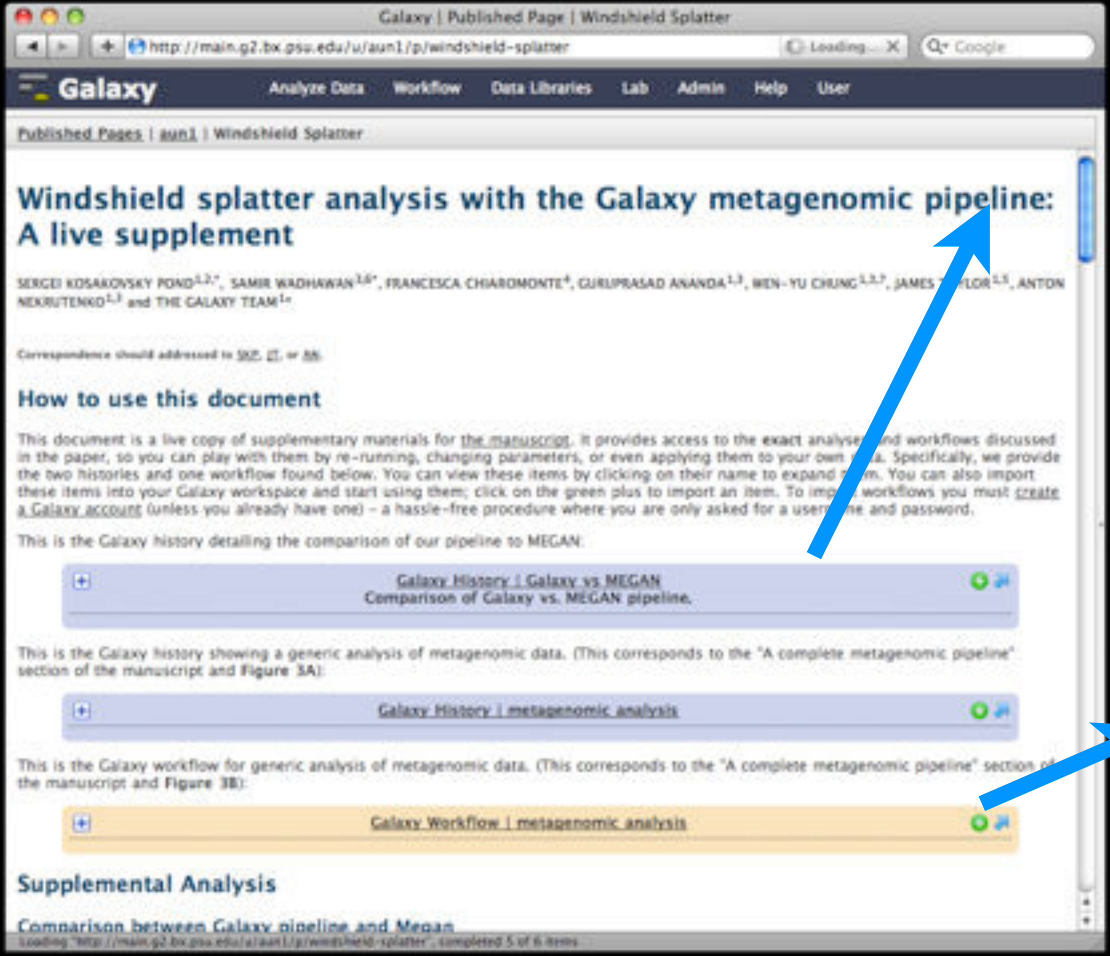
Supplemental Analysis

Comparison between Galaxy pipeline and Megan

Loading "http://main.g2.bx.psu.edu/u/aun1/p/windshield-splatter", completed 5 of 6 items

- All analysis components (datasets, histories, workflows) can be *shared* among Galaxy users and *published*
- Pages and annotation allow analysis to be augmented with textual content and provided in the form of an integrated document

Sharing and publishing


Galaxy | Published Page | Windshield Splatter
http://main.g2.bx.psu.edu/u/aun1/p/windshield-splatter
Loading...
Google
Galaxy
Analyze Data
Workflow
Data Libraries
Lab
Admin
Help
User

Published Pages | aun1 | Windshield Splatter

Windshield splatter analysis with the Galaxy metagenomic pipeline: A live supplement

SERGEI KOSAKOVSKY POND[1,2,*], SAMIR WADHAWAN[1,6*], FRANCESCA CHIAROMONTE[4], GURUPRASAD ANANDA[1,3], WEN-YU CHUNG[1,3,7], JAMES TAYLOR[1,5], ANTON NEKRUTENKO[1,3] and THE GALAXY TEAM[1*]

Correspondence should addressed to SKP, JT, or AN.

How to use this document

This document is a live copy of supplementary materials for the manuscript. It provides access to the **exact** analyses and workflows discussed in the paper, so you can play with them by re-running, changing parameters, or even applying them to your own data. Specifically, we provide the two histories and one workflow found below. You can view these items by clicking on their name to expand them. You can also import these items into your Galaxy workspace and start using them; click on the green plus to import an item. To import workflows you must create a Galaxy account (unless you already have one) – a hassle-free procedure where you are only asked for a username and password.

This is the Galaxy history detailing the comparison of our pipeline to MEGAN:

Galaxy History | Galaxy vs MEGAN
Comparison of Galaxy vs. MEGAN pipeline.

This is the Galaxy history showing a generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3A**):

Galaxy History | metagenomic analysis

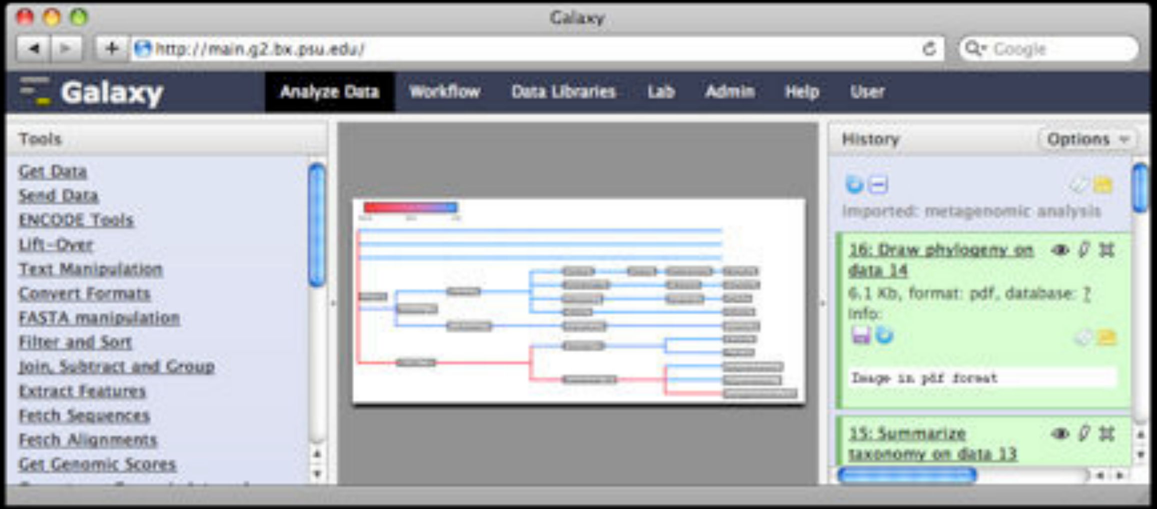
This is the Galaxy workflow for generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3B**):

Galaxy Workflow | metagenomic analysis

Supplemental Analysis

Comparison between Galaxy pipeline and Megan

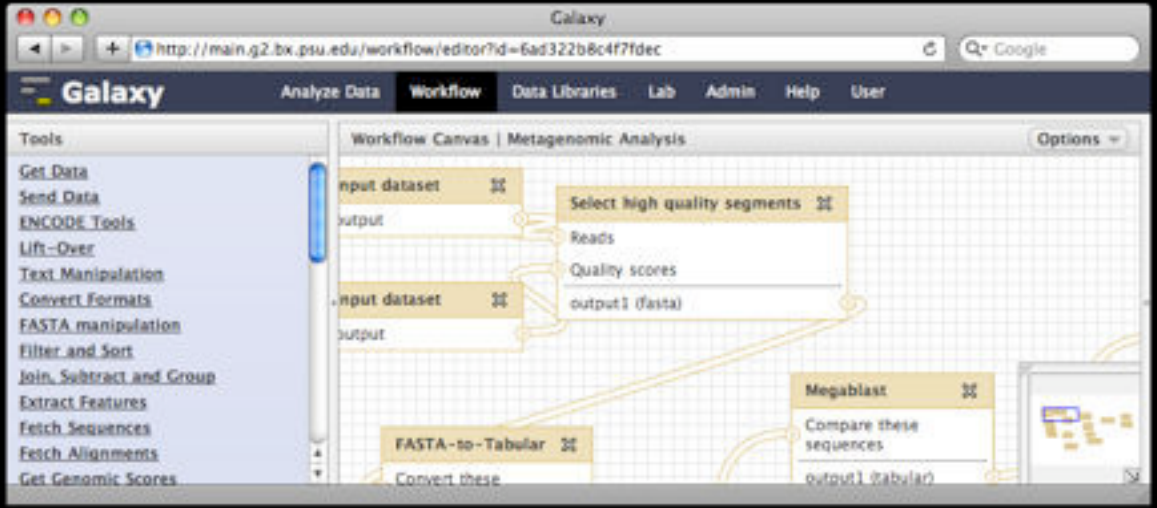
Loading "http://main.g2.bx.psu.edu/u/aun1/p/windshield-splatter", completed 5 of 6 items

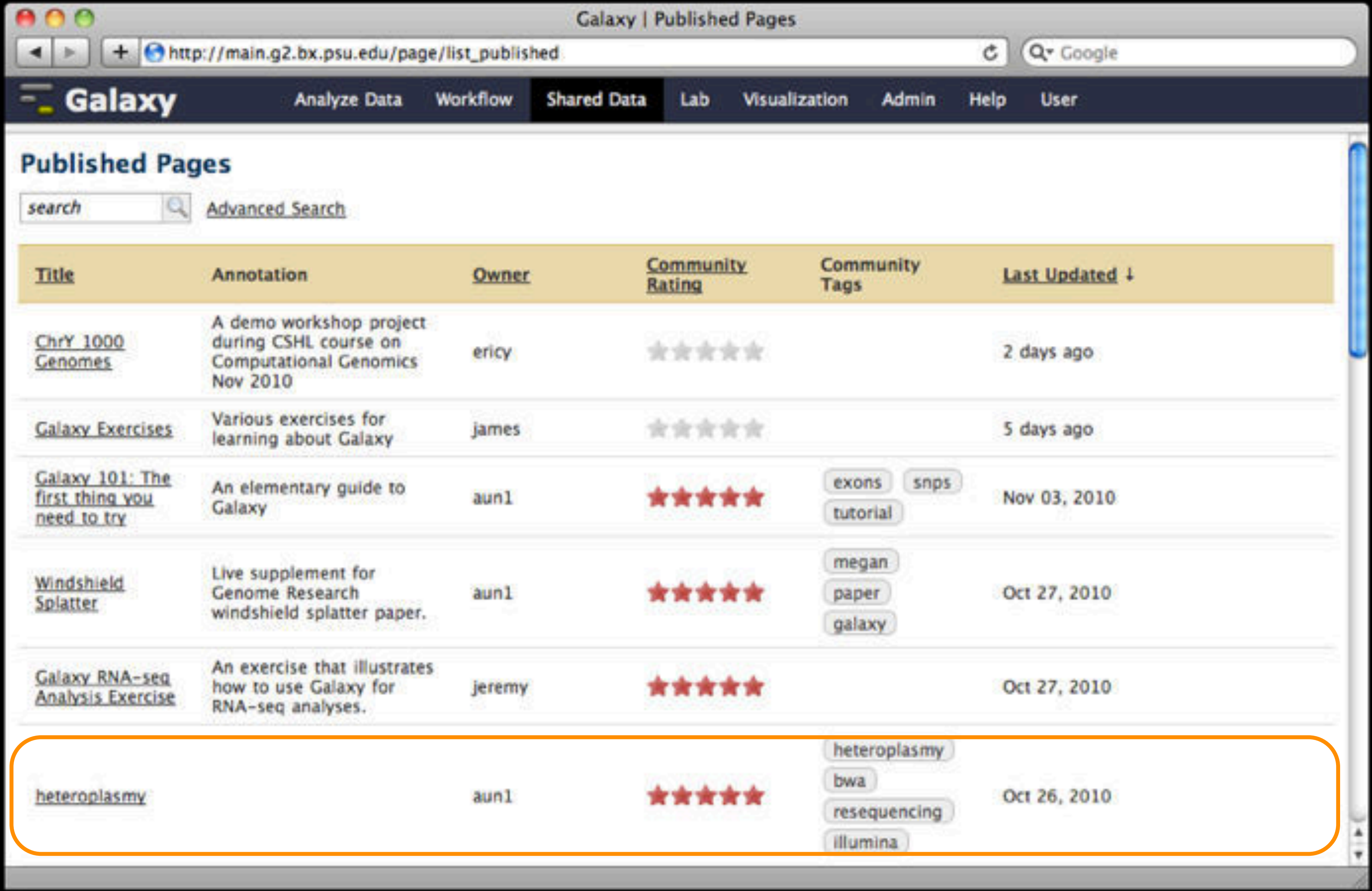


imported: metagenomic analysis

16: Draw phylogeny on data 14
6.1 Kb, format: pdf, database: ?
Info:
Image in pdf format

15: Summarize taxonomy on data 13

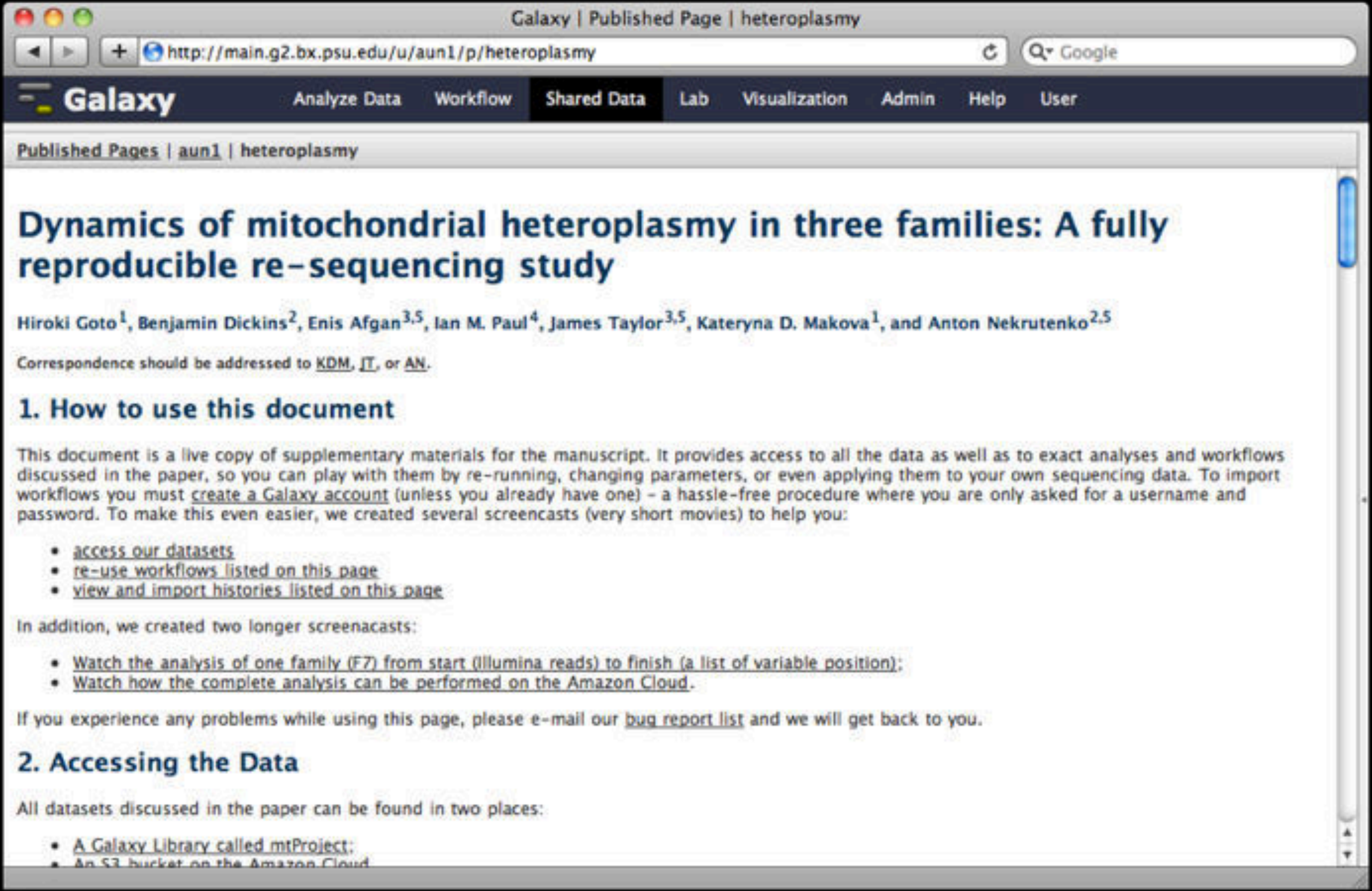




Published Pages

search Advanced Search

Title	Annotation	Owner	Community Rating	Community Tags	Last Updated ↓
ChrY 1000 Genomes	A demo workshop project during CSHL course on Computational Genomics Nov 2010	ericy	☆☆☆☆☆		2 days ago
Galaxy Exercises	Various exercises for learning about Galaxy	james	☆☆☆☆☆		5 days ago
Galaxy 101: The first thing you need to try	An elementary guide to Galaxy	aun1	★★★★★	exons snps tutorial	Nov 03, 2010
Windshield Splatter	Live supplement for Genome Research windshield splatter paper.	aun1	★★★★★	megan paper galaxy	Oct 27, 2010
Galaxy RNA-seq Analysis Exercise	An exercise that illustrates how to use Galaxy for RNA-seq analyses.	jeremy	★★★★★		Oct 27, 2010
heteroplasmy		aun1	★★★★★	heteroplasmy bwa resequencing illumina	Oct 26, 2010


Galaxy | Published Page | heteroplasmy
http://main.g2.bx.psu.edu/u/aun1/p/heteroplasmy
Google
Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
Published Pages | aun1 | heteroplasmy

Dynamics of mitochondrial heteroplasmy in three families: A fully reproducible re-sequencing study

Hiroki Goto[1], Benjamin Dickins[2], Enis Afgan[3,5], Ian M. Paul[4], James Taylor[3,5], Kateryna D. Makova[1], and Anton Nekrutenko[2,5]

Correspondence should be addressed to KDM, JT, or AN.

1. How to use this document

This document is a live copy of supplementary materials for the manuscript. It provides access to all the data as well as to exact analyses and workflows discussed in the paper, so you can play with them by re-running, changing parameters, or even applying them to your own sequencing data. To import workflows you must create a Galaxy account (unless you already have one) - a hassle-free procedure where you are only asked for a username and password. To make this even easier, we created several screencasts (very short movies) to help you:

- access our datasets
- re-use workflows listed on this page
- view and import histories listed on this page

In addition, we created two longer screenacasts:

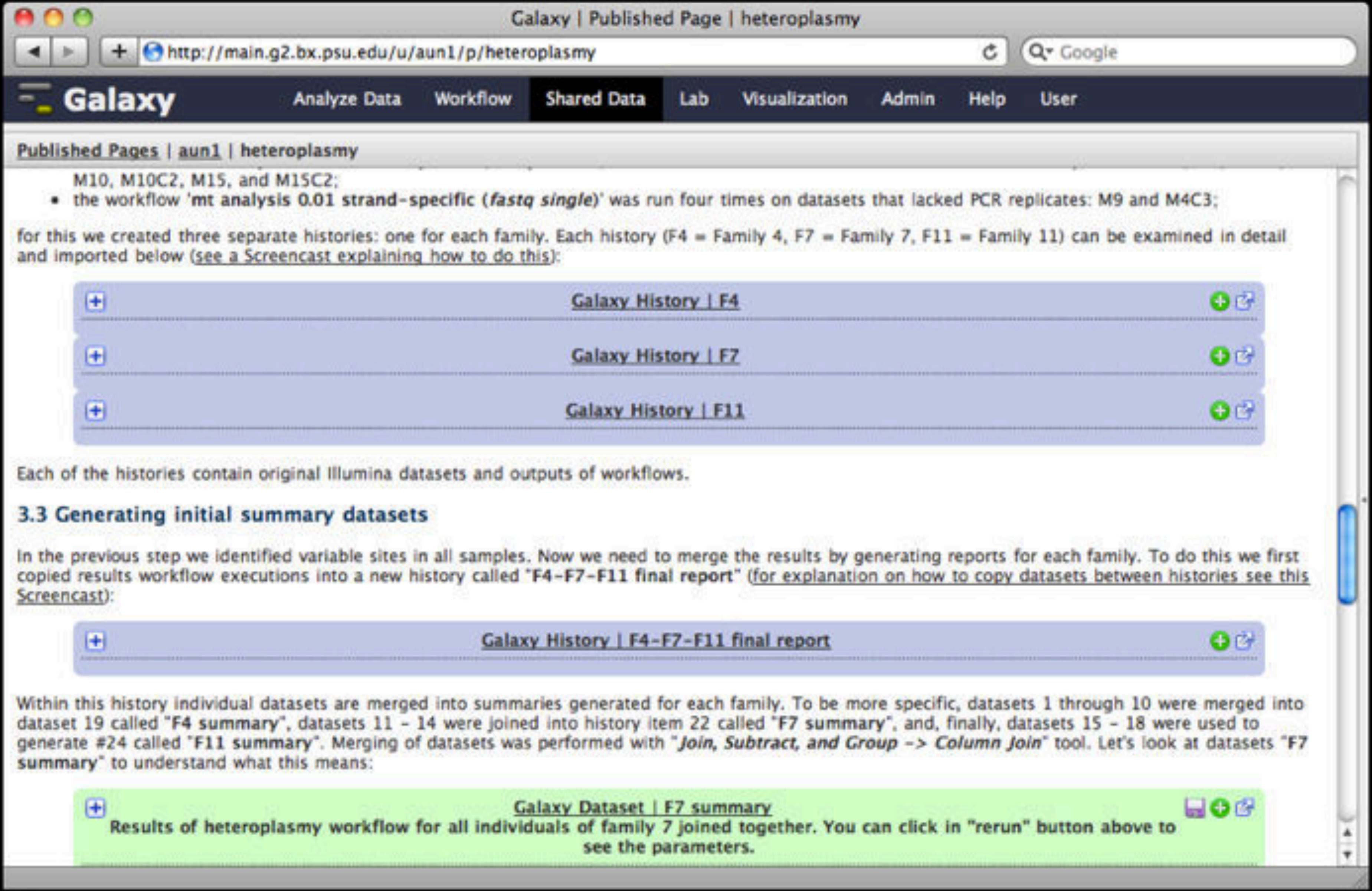
- Watch the analysis of one family (F7) from start (Illumina reads) to finish (a list of variable position);
- Watch how the complete analysis can be performed on the Amazon Cloud.

If you experience any problems while using this page, please e-mail our bug report list and we will get back to you.

2. Accessing the Data

All datasets discussed in the paper can be found in two places:

- A Galaxy Library called mtProject;
- An S3 bucket on the Amazon Cloud


Galaxy | Published Page | heteroplasmy
http://main.g2.bx.psu.edu/u/aun1/p/heteroplasmy
Google
Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
Published Pages | aun1 | heteroplasmy

M10, M10C2, M15, and M15C2;

- the workflow '**mt analysis 0.01 strand-specific (*fastq single*)**' was run four times on datasets that lacked PCR replicates: M9 and M4C3;

for this we created three separate histories: one for each family. Each history (F4 = Family 4, F7 = Family 7, F11 = Family 11) can be examined in detail and imported below (see a Screencast explaining how to do this):

Galaxy History | F4

Galaxy History | F7

Galaxy History | F11

Each of the histories contain original Illumina datasets and outputs of workflows.

3.3 Generating initial summary datasets

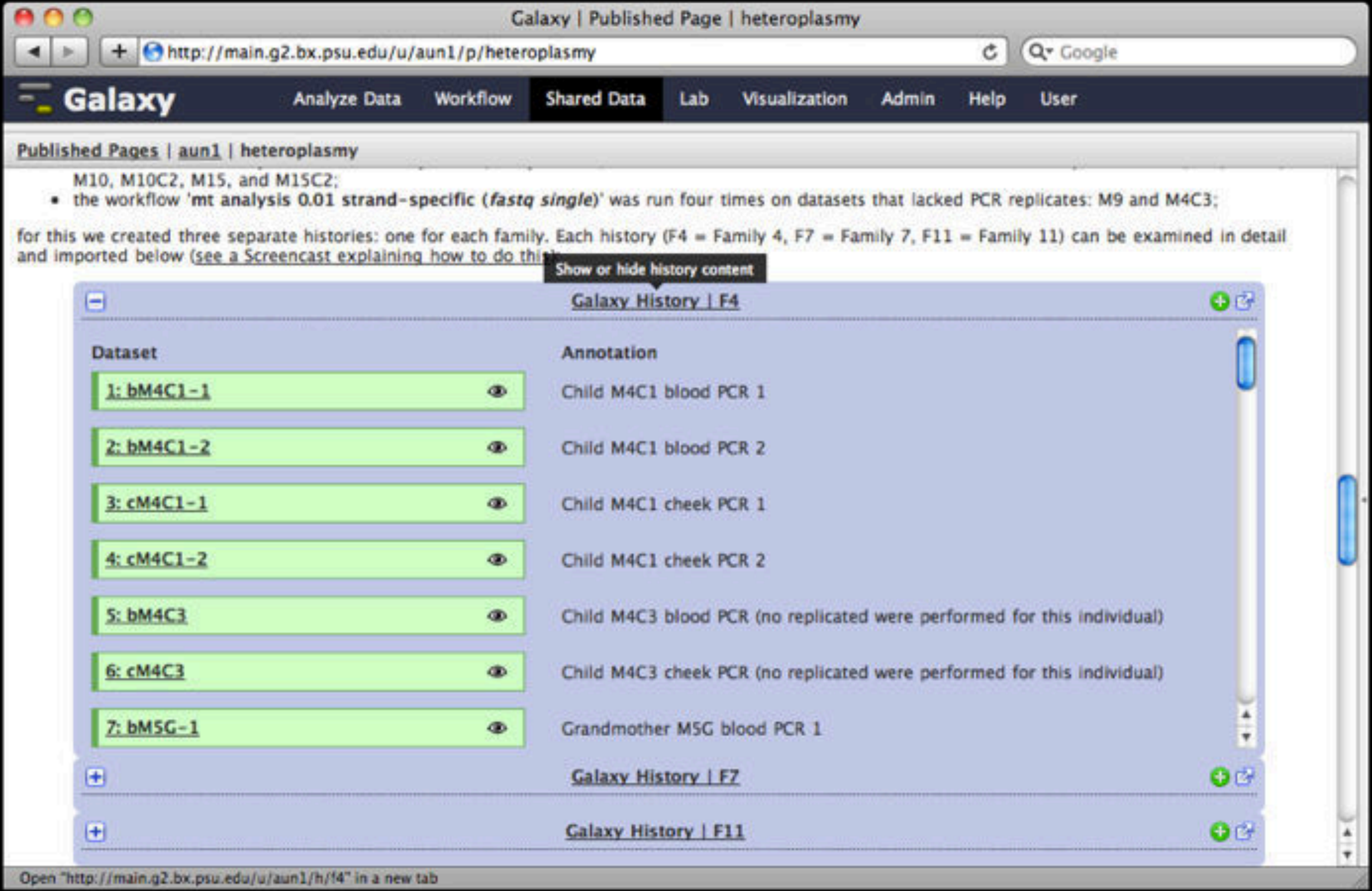
In the previous step we identified variable sites in all samples. Now we need to merge the results by generating reports for each family. To do this we first copied results workflow executions into a new history called "**F4-F7-F11 final report**" (for explanation on how to copy datasets between histories see this Screencast):

Galaxy History | F4-F7-F11 final report

Within this history individual datasets are merged into summaries generated for each family. To be more specific, datasets 1 through 10 were merged into dataset 19 called "**F4 summary**", datasets 11 - 14 were joined into history item 22 called "**F7 summary**", and, finally, datasets 15 - 18 were used to generate #24 called "**F11 summary**". Merging of datasets was performed with "***Join, Subtract, and Group -> Column Join***" tool. Let's look at datasets "**F7 summary**" to understand what this means:

Galaxy Dataset | F7 summary

Results of heteroplasmy workflow for all individuals of family 7 joined together. You can click in "rerun" button above to see the parameters.


Galaxy | Published Page | heteroplasmy
http://main.g2.bx.psu.edu/u/aun1/p/heteroplasmy
Google
Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
Published Pages | aun1 | heteroplasmy
Show or hide history content

M10, M10C2, M15, and M15C2;

- the workflow '**mt analysis 0.01 strand-specific (*fastq single*)**' was run four times on datasets that lacked PCR replicates: M9 and M4C3;

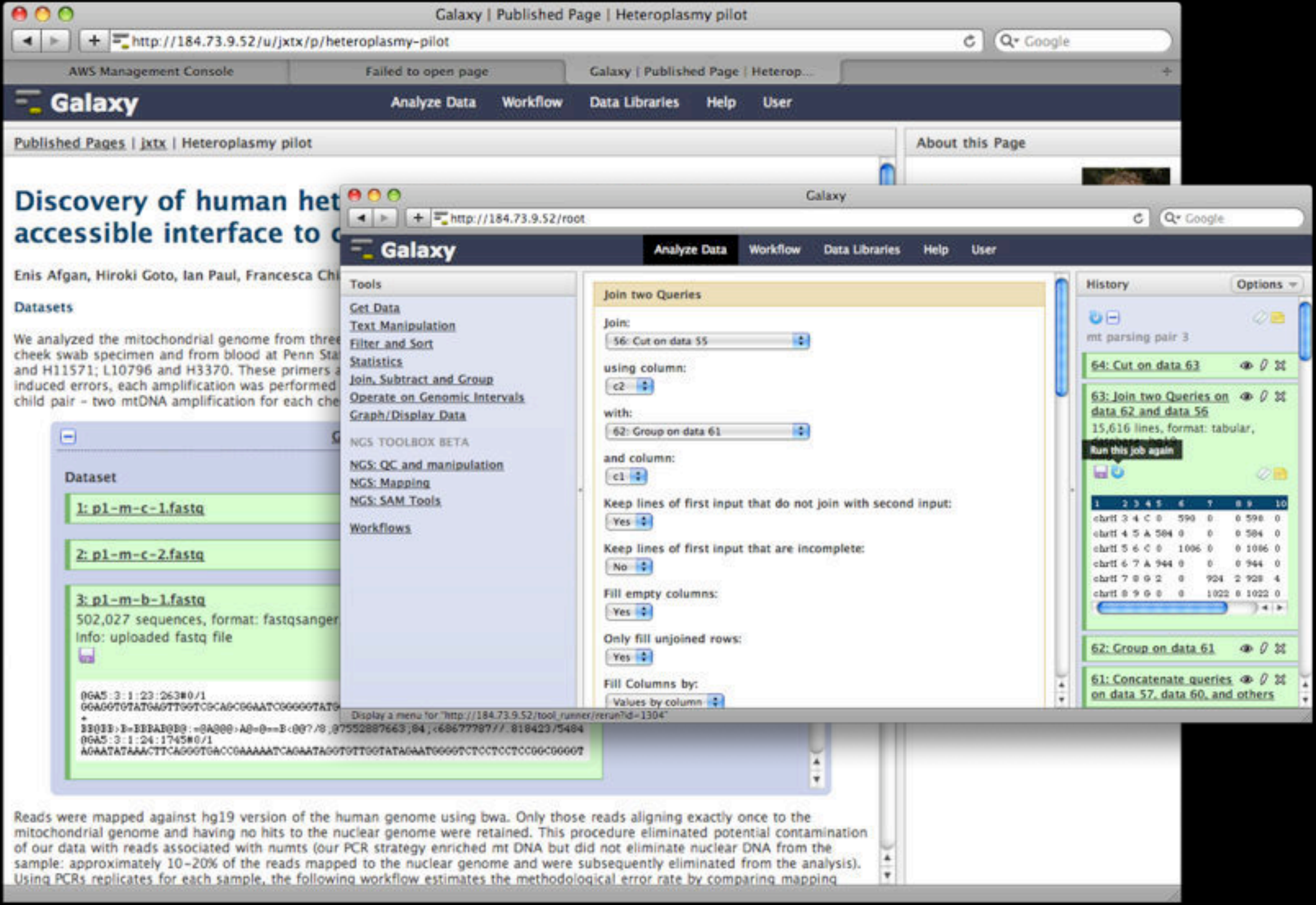
for this we created three separate histories: one for each family. Each history (F4 = Family 4, F7 = Family 7, F11 = Family 11) can be examined in detail and imported below (see a Screencast explaining how to do this)

Galaxy History | F4

Dataset	Annotation
1: bM4C1-1	Child M4C1 blood PCR 1
2: bM4C1-2	Child M4C1 blood PCR 2
3: cM4C1-1	Child M4C1 cheek PCR 1
4: cM4C1-2	Child M4C1 cheek PCR 2
5: bM4C3	Child M4C3 blood PCR (no replicated were performed for this individual)
6: cM4C3	Child M4C3 cheek PCR (no replicated were performed for this individual)
7: bM5G-1	Grandmother M5G blood PCR 1

Galaxy History | F7

Galaxy History | F11



Galaxy | Published Page | Heteroplasmy pilot
http://184.73.9.52/u/jxtx/p/heteroplasmy-pilot
Google
AWS Management Console
Failed to open page
Galaxy | Published Page | Heterop...
Galaxy
Analyze Data
Workflow
Data Libraries
Help
User
Published Pages | jxtx | Heteroplasmy pilot
About this Page

Discovery of human het
accessible interface to c

Enis Afgan, Hiroki Goto, Ian Paul, Francesca Chi

Datasets

We analyzed the mitochondrial genome from three
cheek swab specimen and from blood at Penn Sta
and H11571; L10796 and H3370. These primers a
induced errors, each amplification was performed
child pair - two mtDNA amplification for each che

Dataset

1: p1-m-c-1.fastq

2: p1-m-c-2.fastq

3: p1-m-b-1.fastq

502,027 sequences, format: fastqsanger

Info: uploaded fastq file

```
@GA5:3:1:23:263#0/1
GGAGGTGTATGAGTTGGTCGCAGCGGAATCGGGGGTATG
+
BB@BB>B=BBBAB@B@:=@A@@@>A@=@==B<@@?/8;@7552887663;84;<68677787//.818423/5484
@GA5:3:1:24:1745#0/1
AGAATATAAACTTCAGGGTGACCGAAAAATCAGAATAGGTGTTGGTATAGAATGGGGTCTCCTCCTCCGGCGGGGT
```

Reads were mapped against hg19 version of the human genome using bwa. Only those reads aligning exactly once to the mitochondrial genome and having no hits to the nuclear genome were retained. This procedure eliminated potential contamination of our data with reads associated with numts (our PCR strategy enriched mt DNA but did not eliminate nuclear DNA from the sample: approximately 10-20% of the reads mapped to the nuclear genome and were subsequently eliminated from the analysis). Using PCRs replicates for each sample, the following workflow estimates the methodological error rate by comparing mapping

Join two Queries

Join: 56: Cut on data 55

using column: c2

with: 62: Group on data 61

and column: c1

Keep lines of first input that do not join with second input: Yes

Keep lines of first input that are incomplete: No

Fill empty columns: Yes

Only fill unjoined rows: Yes

Fill Columns by: Values by column

Display a menu for "http://184.73.9.52/tool_runner/rerun?id=1304"

64: Cut on data 63

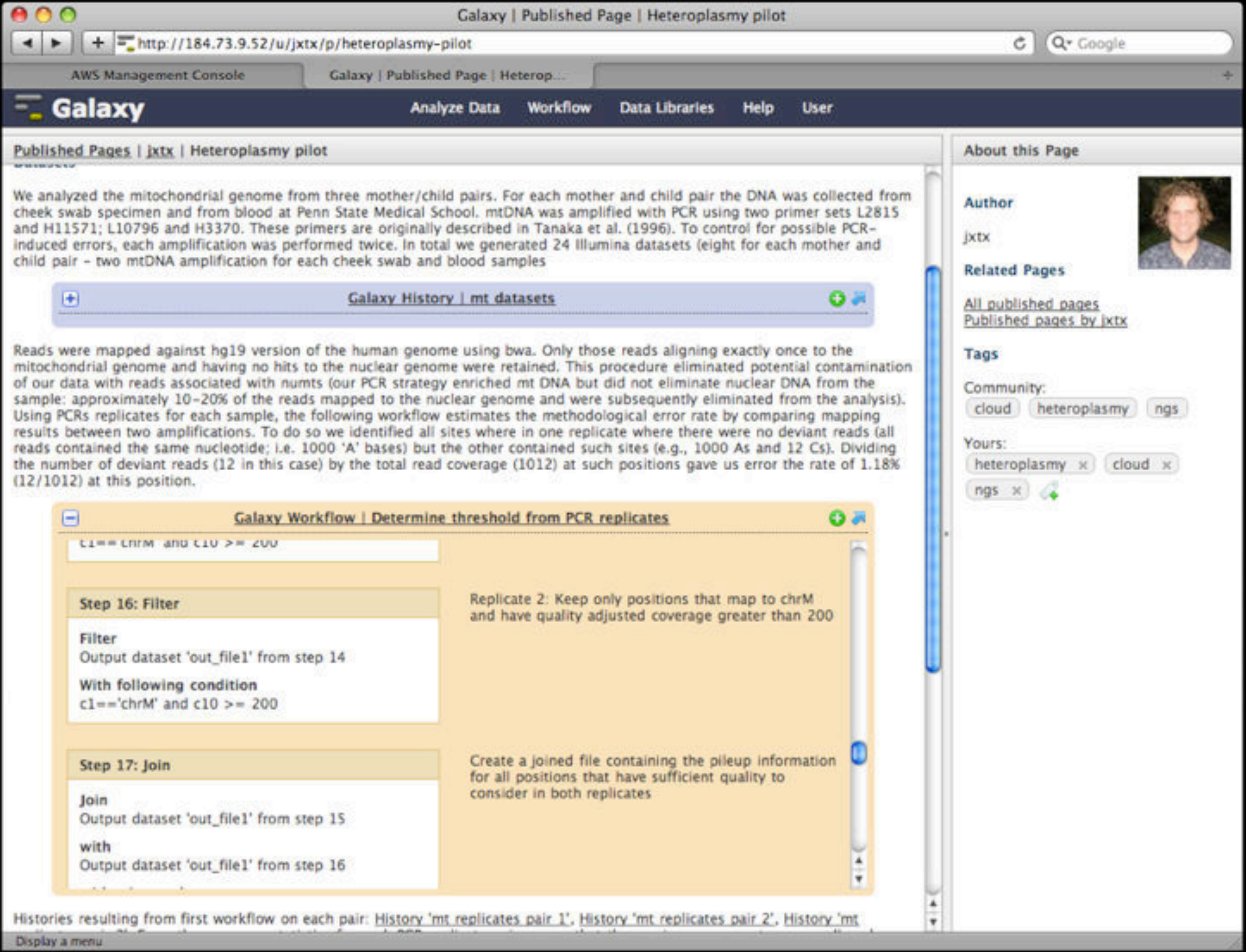
63: Join two Queries on data 62 and data 56

15,616 lines, format: tabular,

1	2	3	4	5	6	7	8	9	10
chrM	3	4	C	0	590	0	0	590	0
chrM	4	5	A	584	0	0	0	584	0
chrM	5	6	C	0	1006	0	0	1006	0
chrM	6	7	A	944	0	0	0	944	0
chrM	7	8	G	2	0	924	2	928	4
chrM	8	9	G	0	0	1022	0	1022	0

62: Group on data 61

61: Concatenate queries on data 57, data 60, and others



Galaxy | Published Page | Heteroplasmy pilot
http://184.73.9.52/u/jxtx/p/heteroplasmy-pilot
Google
AWS Management Console
Galaxy | Published Page | Heterop...
Galaxy
Analyze Data
Workflow
Data Libraries
Help
User
Published Pages | jxtx | Heteroplasmy pilot
About this Page

We analyzed the mitochondrial genome from three mother/child pairs. For each mother and child pair the DNA was collected from cheek swab specimen and from blood at Penn State Medical School. mtDNA was amplified with PCR using two primer sets L2815 and H11571; L10796 and H3370. These primers are originally described in Tanaka et al. (1996). To control for possible PCR-induced errors, each amplification was performed twice. In total we generated 24 Illumina datasets (eight for each mother and child pair - two mtDNA amplification for each cheek swab and blood samples

Galaxy History | mt datasets

Reads were mapped against hg19 version of the human genome using bwa. Only those reads aligning exactly once to the mitochondrial genome and having no hits to the nuclear genome were retained. This procedure eliminated potential contamination of our data with reads associated with numts (our PCR strategy enriched mt DNA but did not eliminate nuclear DNA from the sample: approximately 10-20% of the reads mapped to the nuclear genome and were subsequently eliminated from the analysis). Using PCRs replicates for each sample, the following workflow estimates the methodological error rate by comparing mapping results between two amplifications. To do so we identified all sites where in one replicate where there were no deviant reads (all reads contained the same nucleotide; i.e. 1000 'A' bases) but the other contained such sites (e.g., 1000 As and 12 Cs). Dividing the number of deviant reads (12 in this case) by the total read coverage (1012) at such positions gave us error the rate of 1.18% (12/1012) at this position.

Galaxy Workflow | Determine threshold from PCR replicates

c1=='chrM' and c10 >= 200

Step 16: Filter

Filter
Output dataset 'out_file1' from step 14

With following condition
c1=='chrM' and c10 >= 200

Replicate 2: Keep only positions that map to chrM and have quality adjusted coverage greater than 200

Step 17: Join

Join
Output dataset 'out_file1' from step 15

with
Output dataset 'out_file1' from step 16

Create a joined file containing the pileup information for all positions that have sufficient quality to consider in both replicates

Histories resulting from first workflow on each pair: History 'mt replicates pair 1', History 'mt replicates pair 2', History 'mt

Author

jxtx

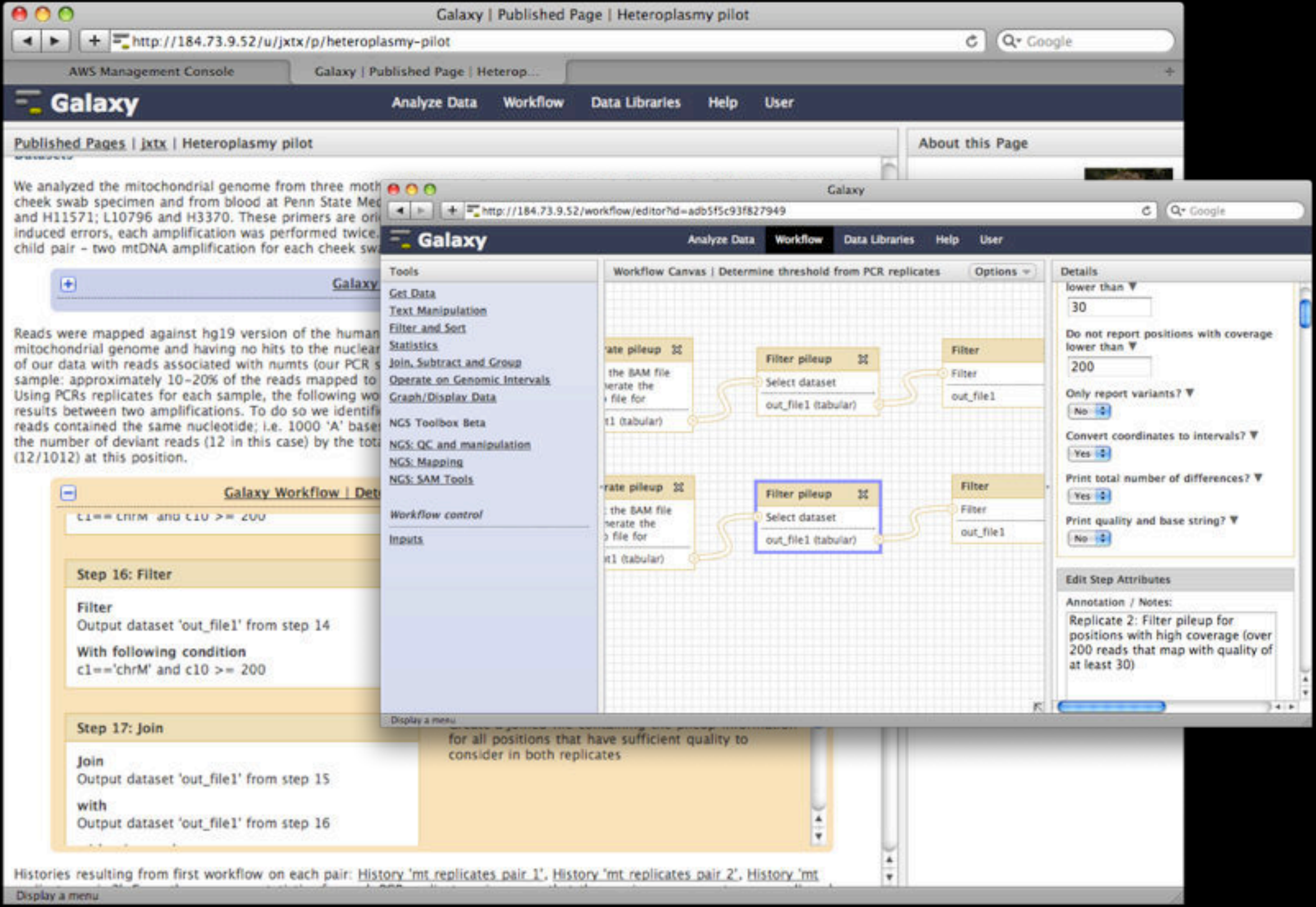
Related Pages

All published pages
Published pages by jxtx

Tags

Community:
cloud heteroplasmy ngs

Yours:
heteroplasmy × cloud × ngs ×



We analyzed the mitochondrial genome from three moth
cheek swab specimen and from blood at Penn State Med
and H11571; L10796 and H3370. These primers are ori
induced errors, each amplification was performed twice.
child pair – two mtDNA amplification for each cheek sw

Galaxy

Reads were mapped against hg19 version of the human
mitochondrial genome and having no hits to the nuclear
of our data with reads associated with numts (our PCR s
sample: approximately 10–20% of the reads mapped to
Using PCRs replicates for each sample, the following wo
results between two amplifications. To do so we identifi
reads contained the same nucleotide; i.e. 1000 'A' base
the number of deviant reads (12 in this case) by the tota
(12/1012) at this position.

Galaxy Workflow | Det

c1=='chrM' and c10 >= 200

Step 16: Filter

Filter
Output dataset 'out_file1' from step 14

With following condition
c1=='chrM' and c10 >= 200

Step 17: Join

Join
Output dataset 'out_file1' from step 15

with
Output dataset 'out_file1' from step 16

for all positions that have sufficient quality to
consider in both replicates

Histories resulting from first workflow on each pair: History 'mt replicates pair 1', History 'mt replicates pair 2', History 'mt

lower than ▼
30

Do not report positions with coverage lower than ▼
200

Only report variants? ▼
No

Convert coordinates to intervals? ▼
Yes

Print total number of differences? ▼
Yes

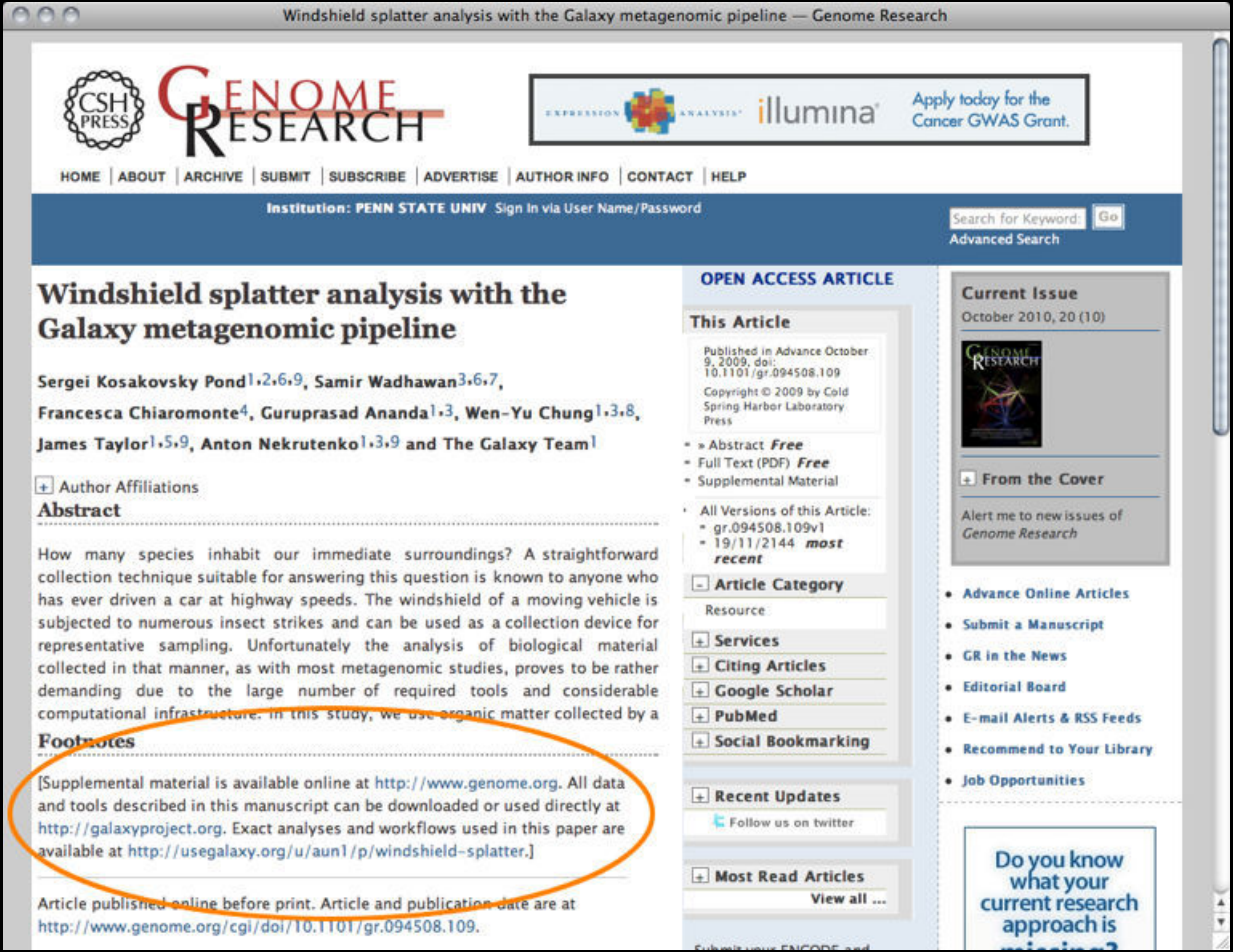
Print quality and base string? ▼
No

Edit Step Attributes

Annotation / Notes:
Replicate 2: Filter pileup for positions with high coverage (over 200 reads that map with quality of at least 30)

The power of Galaxy publishing and sharing

- Galaxy's publishing features facilitate access and reproducibility without any extra leg work
- One click grants access to the *actual analysis* you performed to generate your original results
    - Not just data access: the full pipeline
    - Annotate each step
    - Anyone can import your work and immediately reproduce or build on it



Windshield splatter analysis with the Galaxy metagenomic pipeline — Genome Research

CSH PRESS
GENOME RESEARCH

HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CONTACT | HELP
Institution: PENN STATE UNIV Sign In via User Name/Password
Search for Keyword: Go
Advanced Search

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]

+ Author Affiliations

Abstract

How many species inhabit our immediate surroundings? A straightforward collection technique suitable for answering this question is known to anyone who has ever driven a car at highway speeds. The windshield of a moving vehicle is subjected to numerous insect strikes and can be used as a collection device for representative sampling. Unfortunately the analysis of biological material collected in that manner, as with most metagenomic studies, proves to be rather demanding due to the large number of required tools and considerable computational infrastructure. In this study, we use organic matter collected by a

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.

OPEN ACCESS ARTICLE

This Article

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

- » Abstract *Free*
- Full Text (PDF) *Free*
- Supplemental Material

All Versions of this Article:
- gr.094508.109v1
- 19/11/2144 ***most recent***

□ Article Category

Resource

Current Issue

October 2010, 20 (10)

Alert me to new issues of *Genome Research*

- Advance Online Articles
- Submit a Manuscript
- GR in the News
- Editorial Board
- E-mail Alerts & RSS Feeds
- Recommend to Your Library
- Job Opportunities

Try it now:

http://usegalaxy.org

Develop and deploy:

http://getgalaxy.org