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

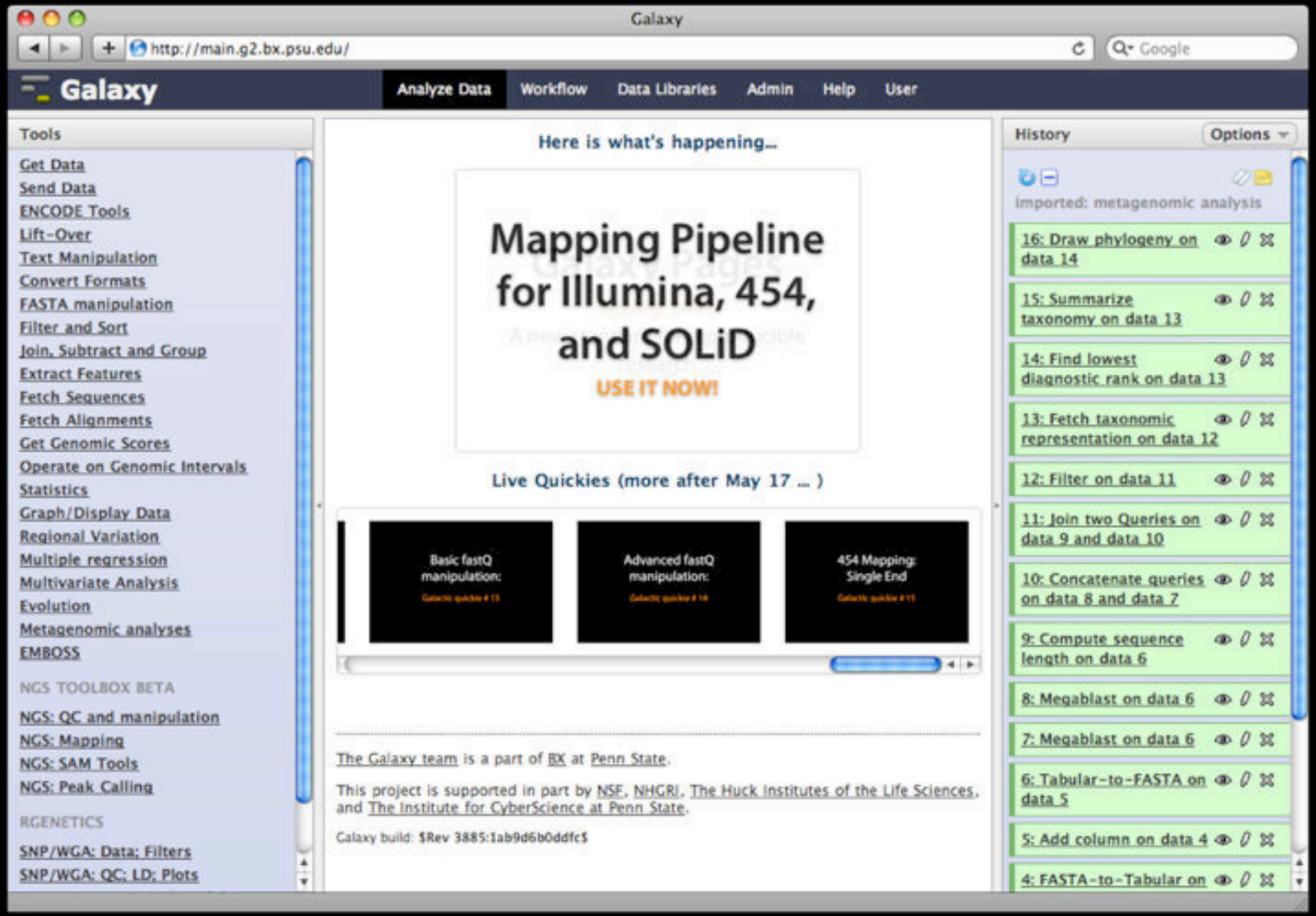
Greg von Kuster

Kanwei Li

Kelly Vincent

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

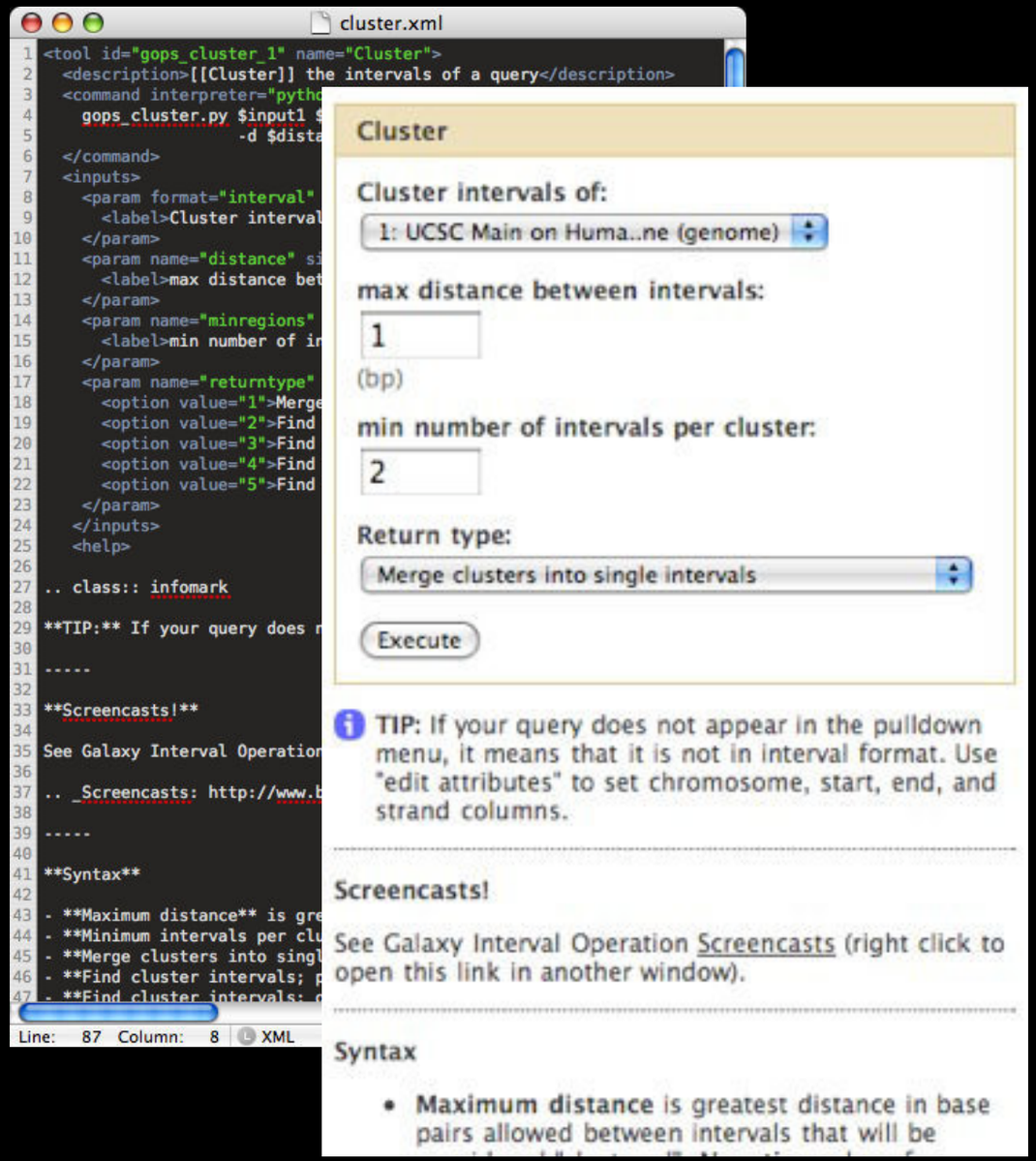
Galaxy: accessible analysis system



What is Galaxy?

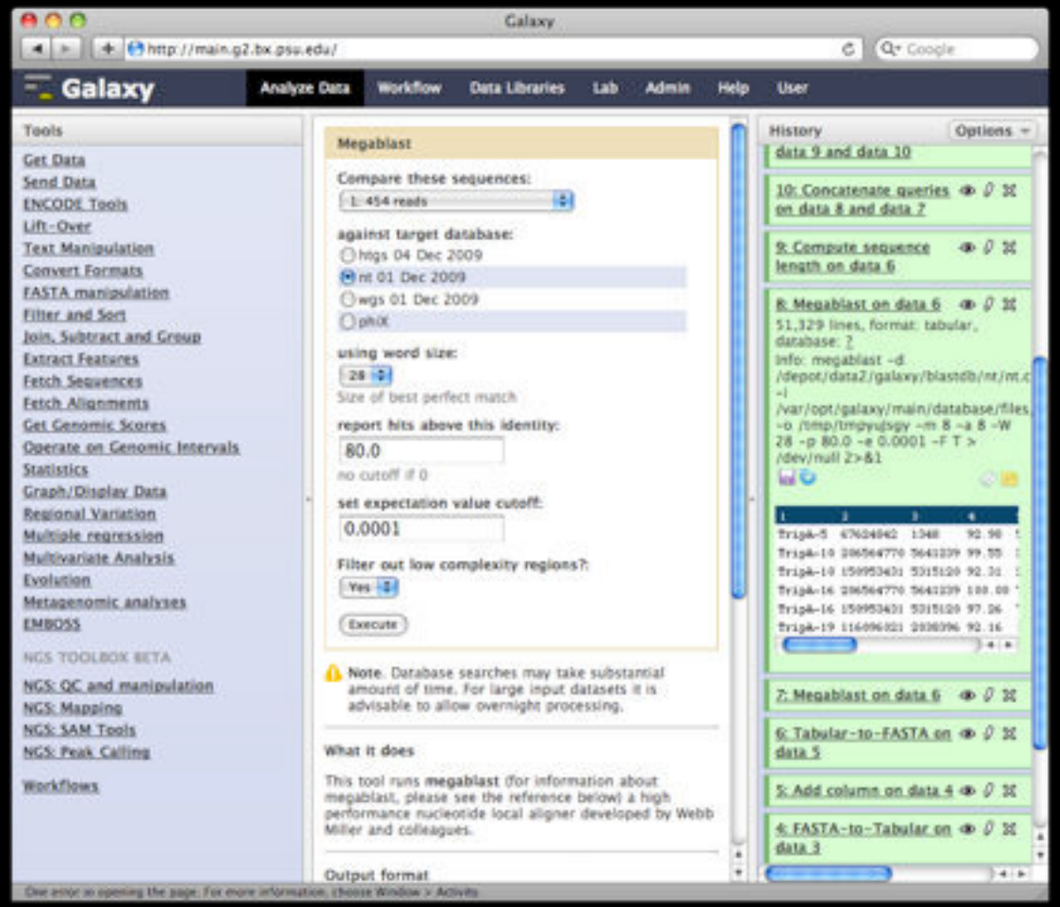
- **A free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- **Open source software** that makes integrating your own tools and data and customizing for your own site simple

Integrating existing tools into a uniform framework



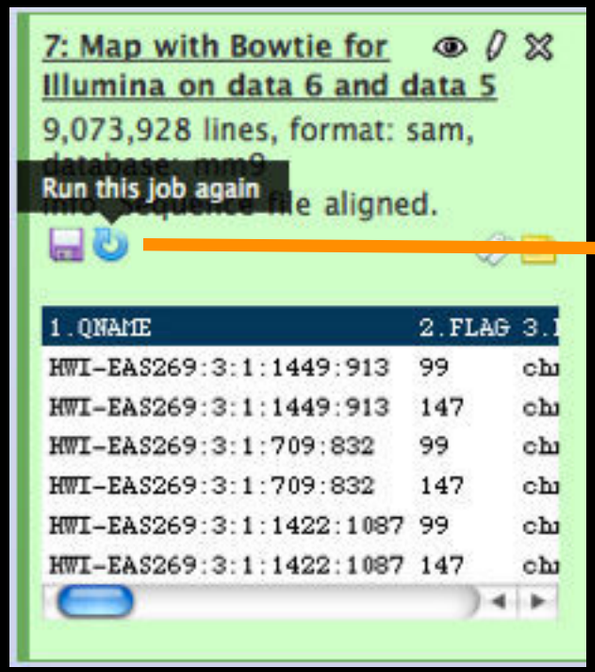
- Defined in terms of an abstract interface (inputs and outputs)
  - In practice, mostly command line tools, a declarative XML description of the interface, how to generate a command line
- Designed to be as easy as possible for tool authors, while still allowing rigorous reasoning

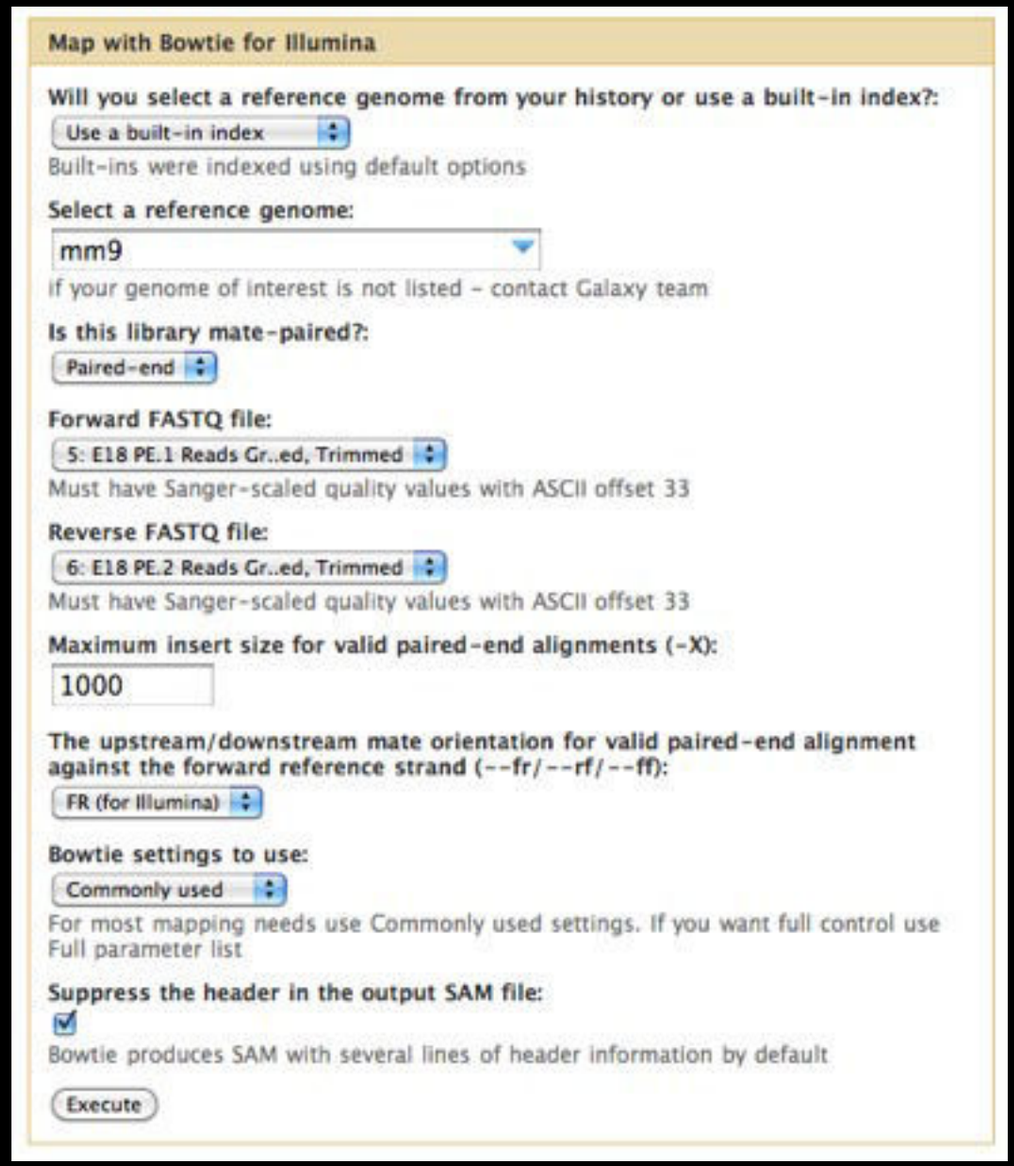
Galaxy analysis interface



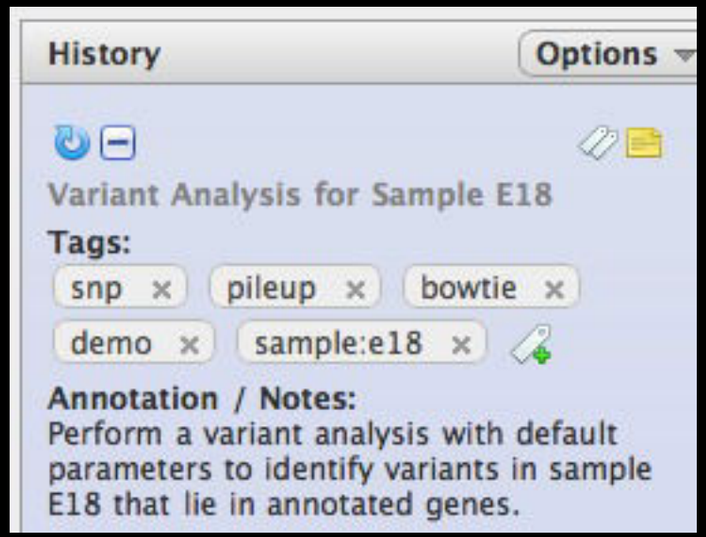
- Consistent tool user interfaces automatically generated
- History system facilitates and tracks multistep analyses

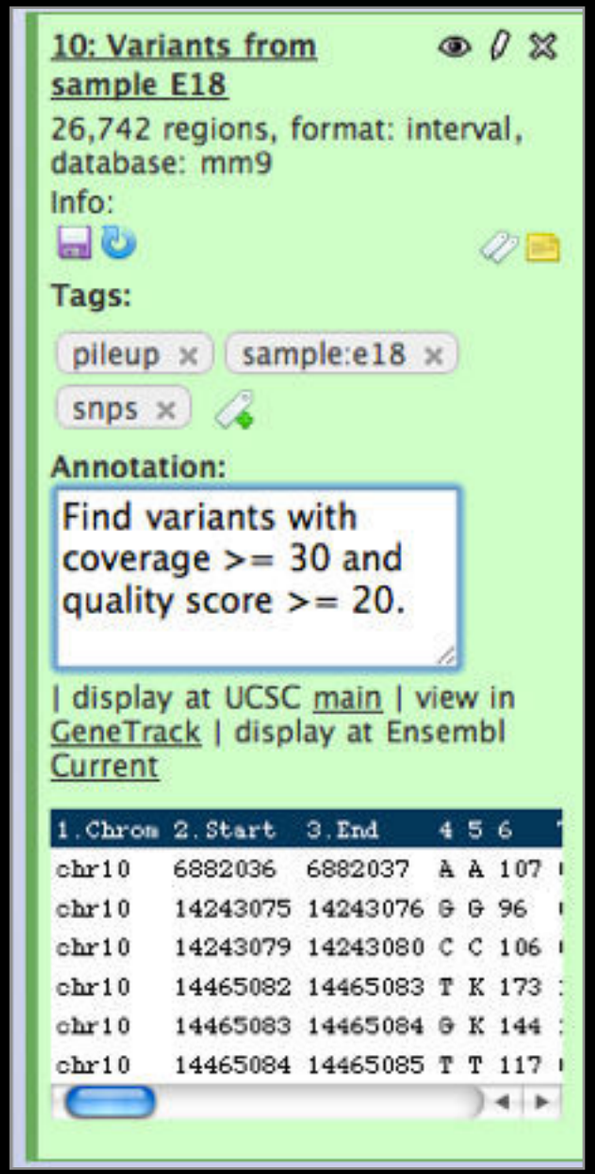
Automatically tracks every step of every analysis



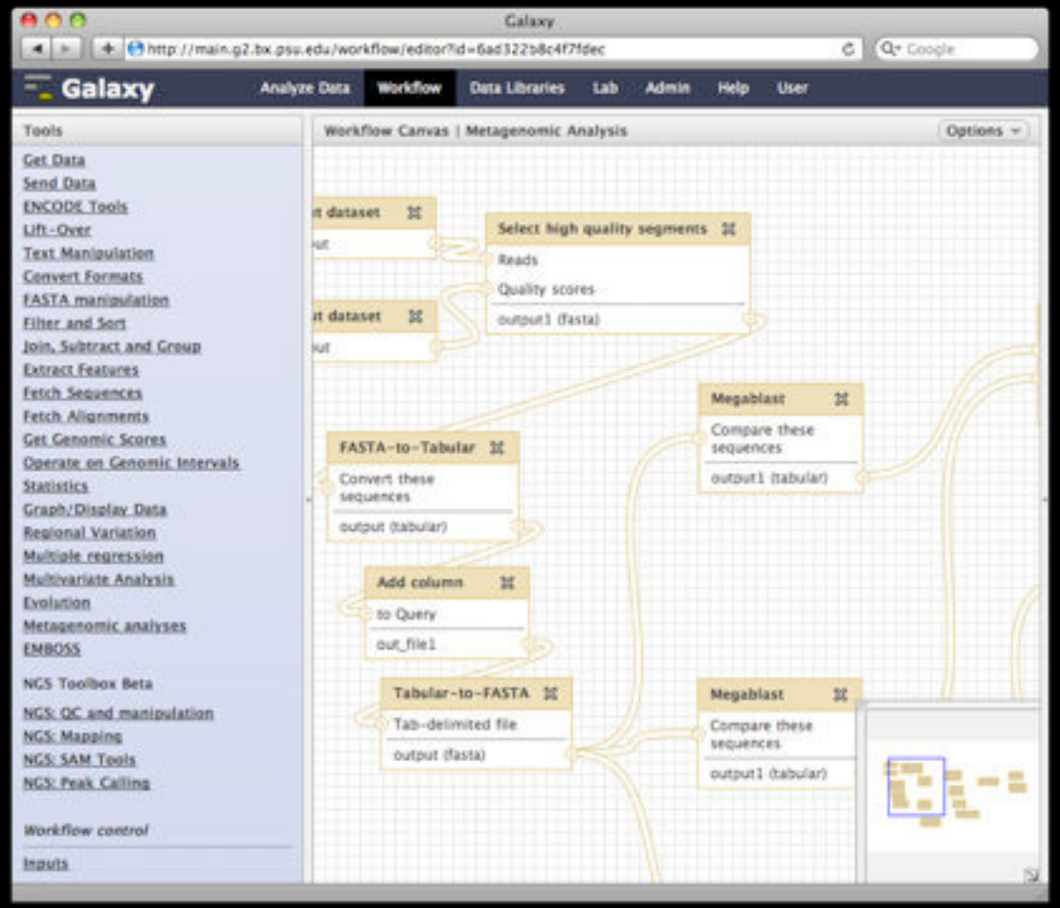


As well as user-generated metadata and annotation...





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

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

NGS: SAM Tools

NGS: Indel Analysis

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

NGS: Peak Calling

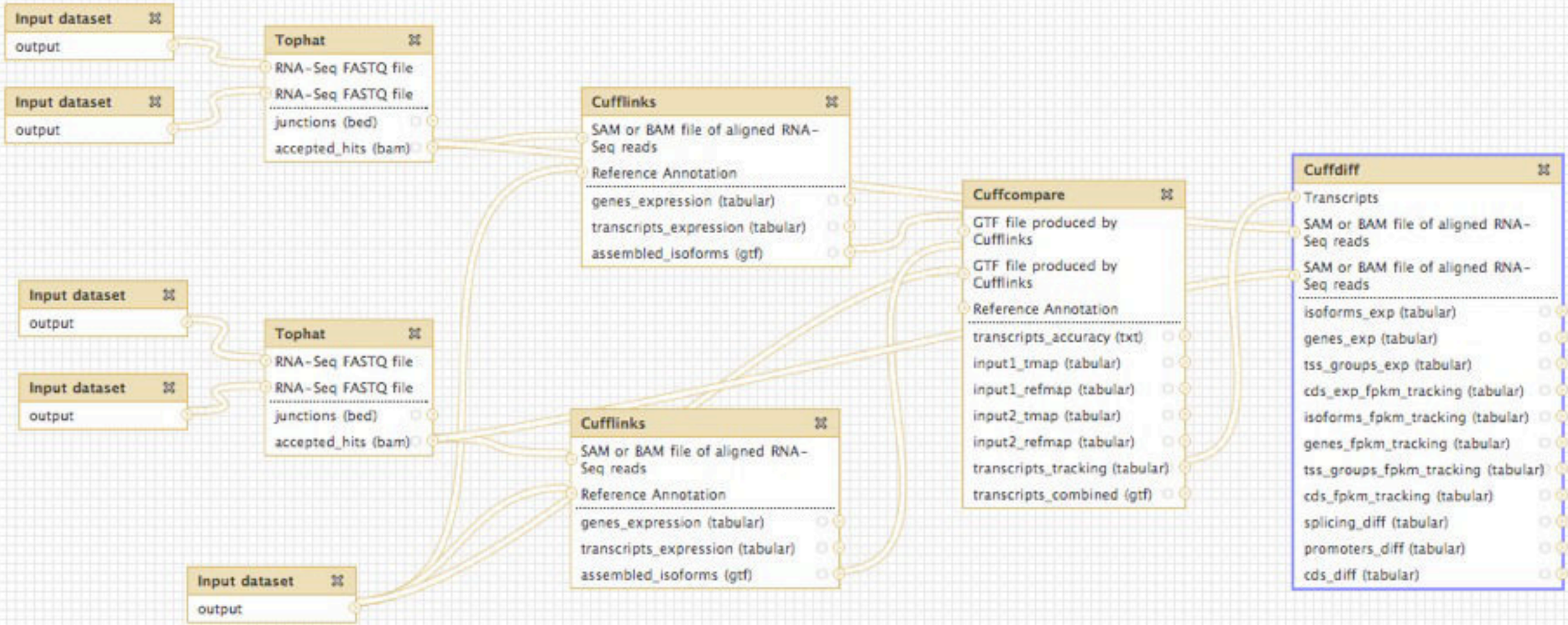
- MACS Model-based Analysis of ChIP-Seq
- GeneTrack indexer on a BED file
- Peak predictor on GeneTrack index

NGS: RNA Analysis

RNA-SEQ

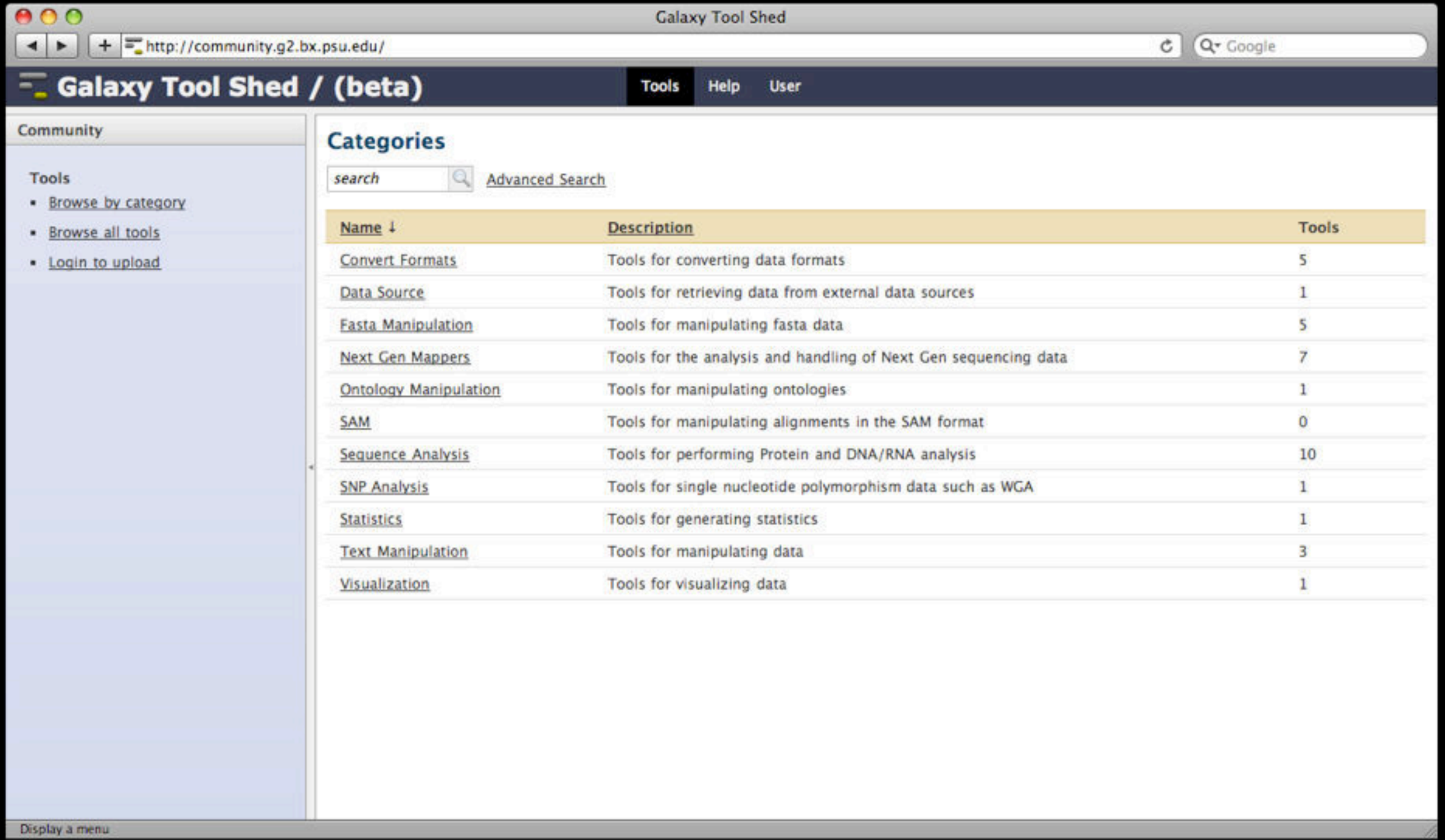
- Tophat Find splice junctions using RNA-seq data
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffdiff find significant changes in transcript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

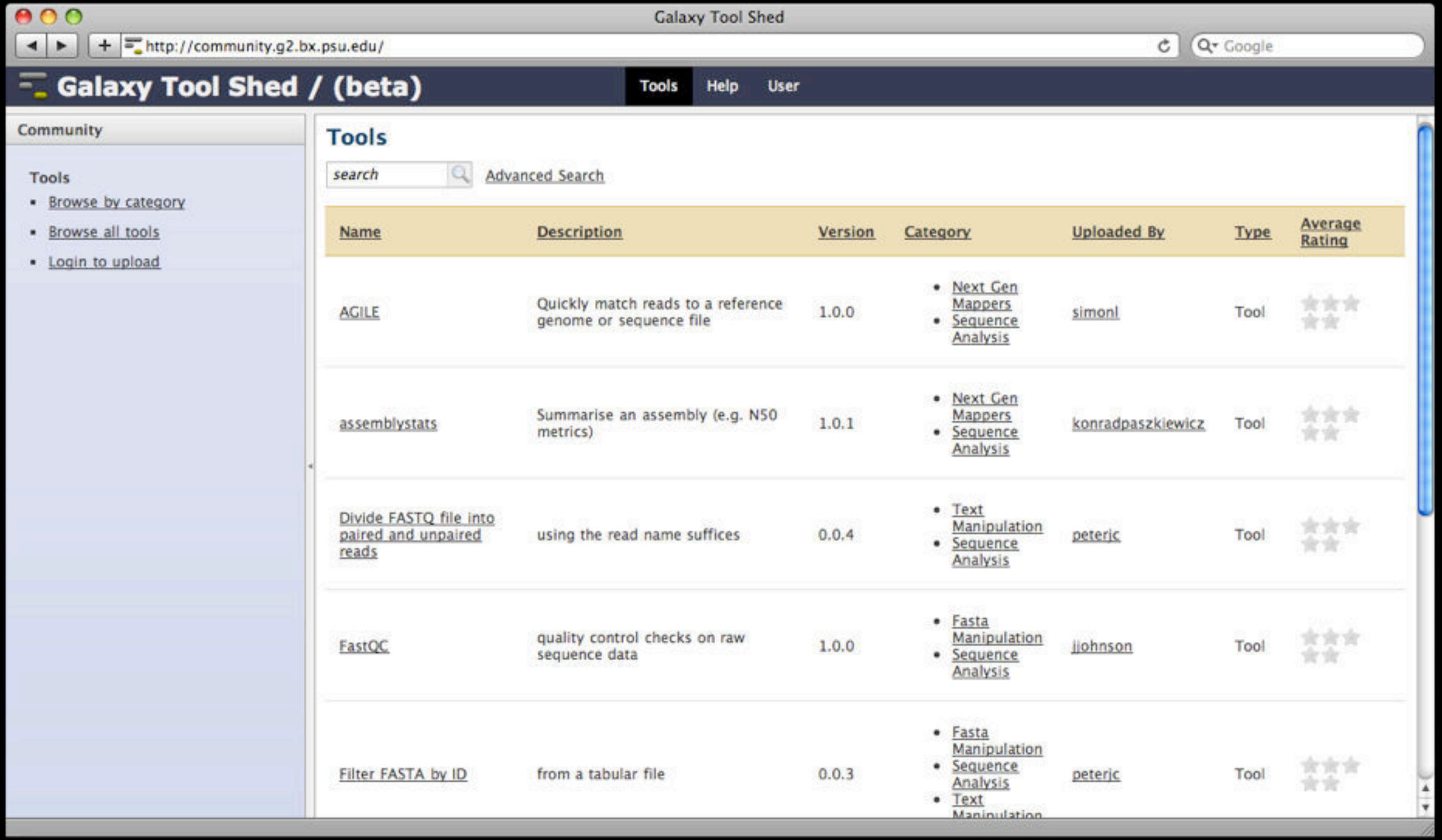
FILTERING



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

Community of tool developers

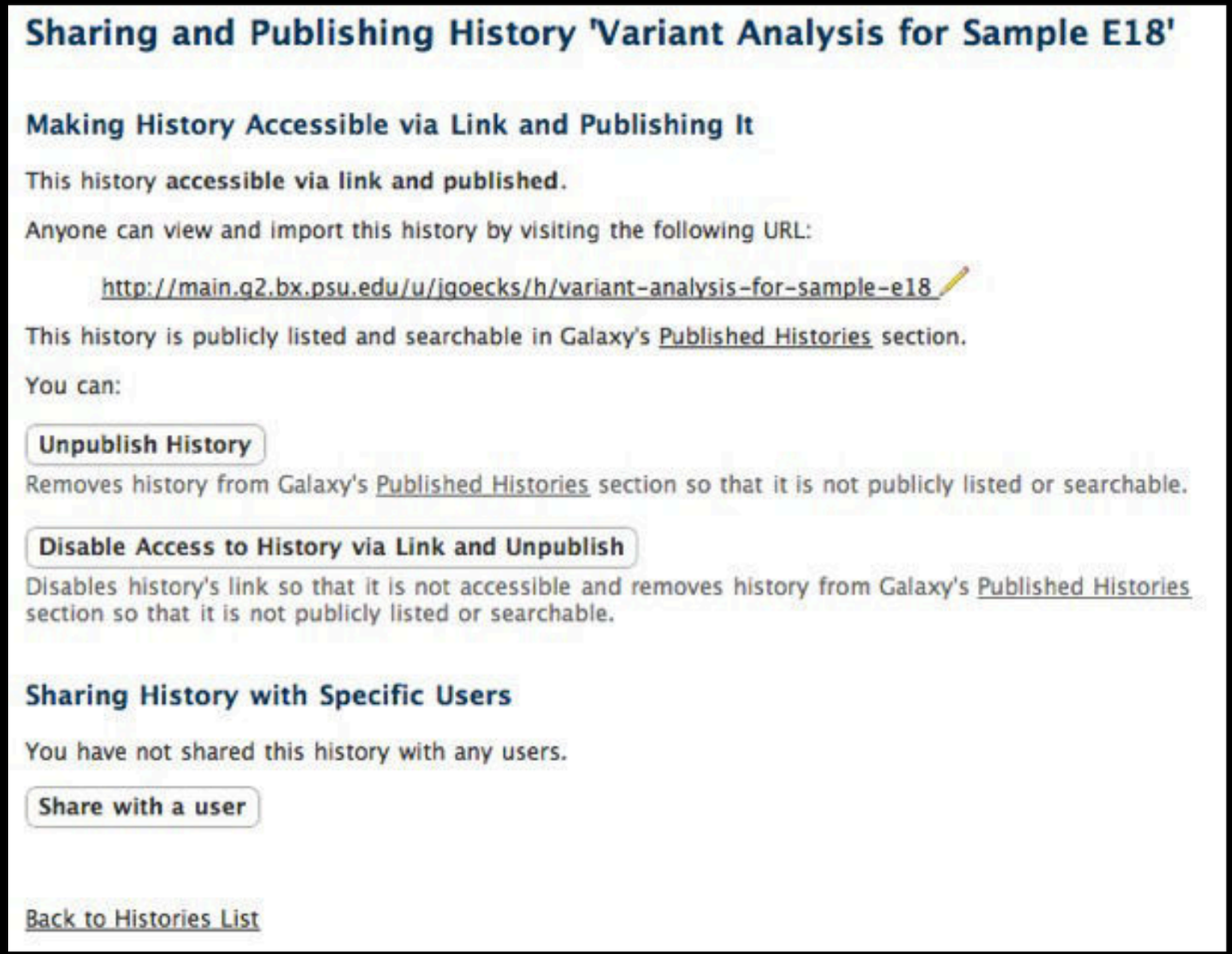

Galaxy Tool Shed
http://community.g2.bx.psu.edu/
Google
Galaxy Tool Shed / (beta)
Tools
Help
User
Community
Tools
Browse by category
Browse all tools
Login to upload
Categories
search
Advanced Search
Name
Description
Tools
Convert Formats
Tools for converting data formats
5
Data Source
Tools for retrieving 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
Display a menu


Galaxy Tool Shed
http://community.g2.bx.psu.edu/
Google
Galaxy Tool Shed / (beta)
Tools
Help
User
Community
Tools
Browse by category
Browse all tools
Login to upload
Tools
search
Advanced Search
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
peterjc
Tool


Galaxy Tool Shed
http://community.g2.bx.psu.edu/
Google
Galaxy Tool Shed / (beta)
Tools
Help
User
Community
Tools
Browse by category
Browse all tools
Login to upload
View Tool
This is the latest approved version of this tool suite
Tool Actions
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
Display a menu

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

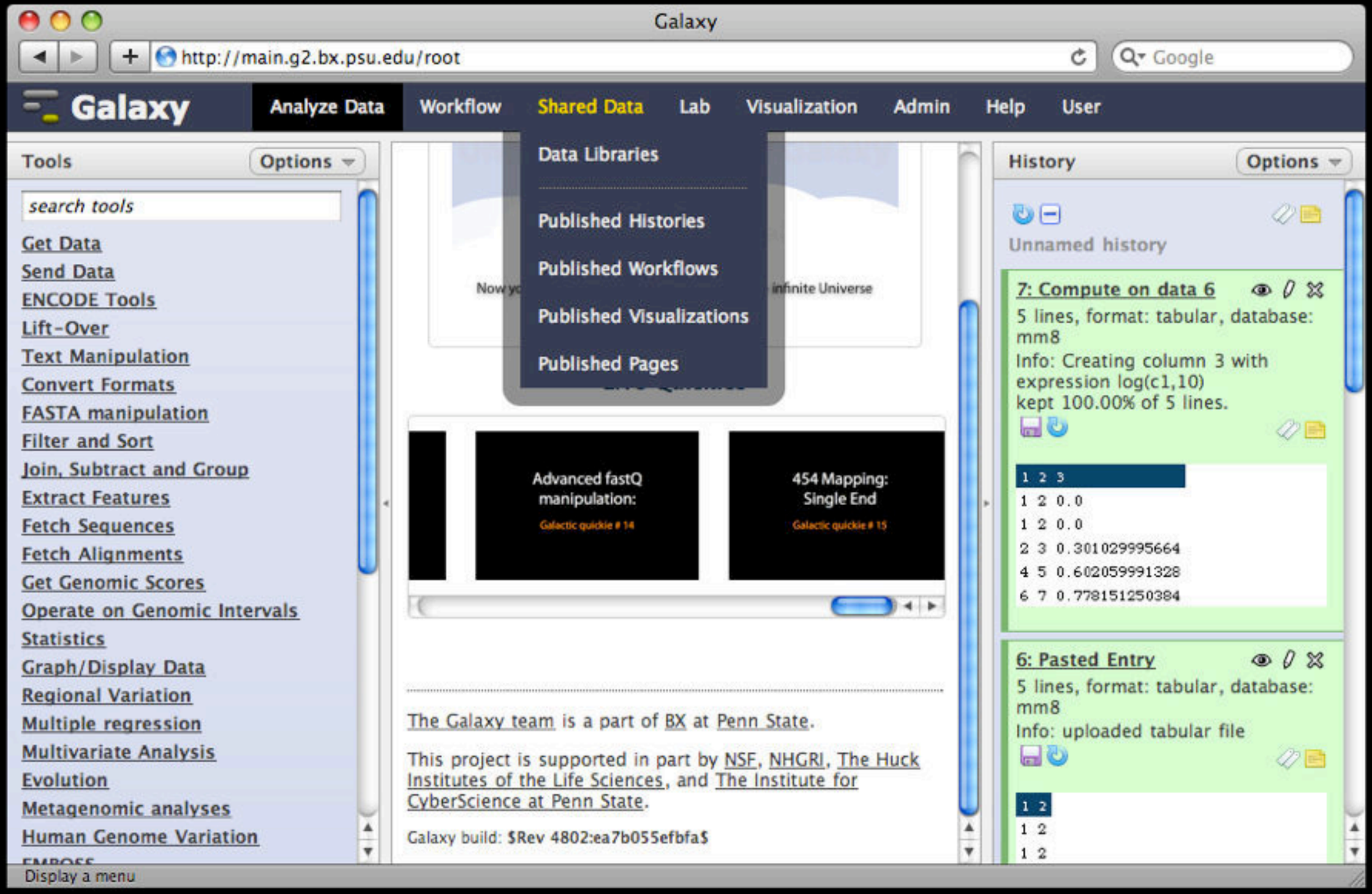
Disables history's link so that it is not accessible and removes history from Galaxy's Published Histories section so that it is not publicly listed or searchable.

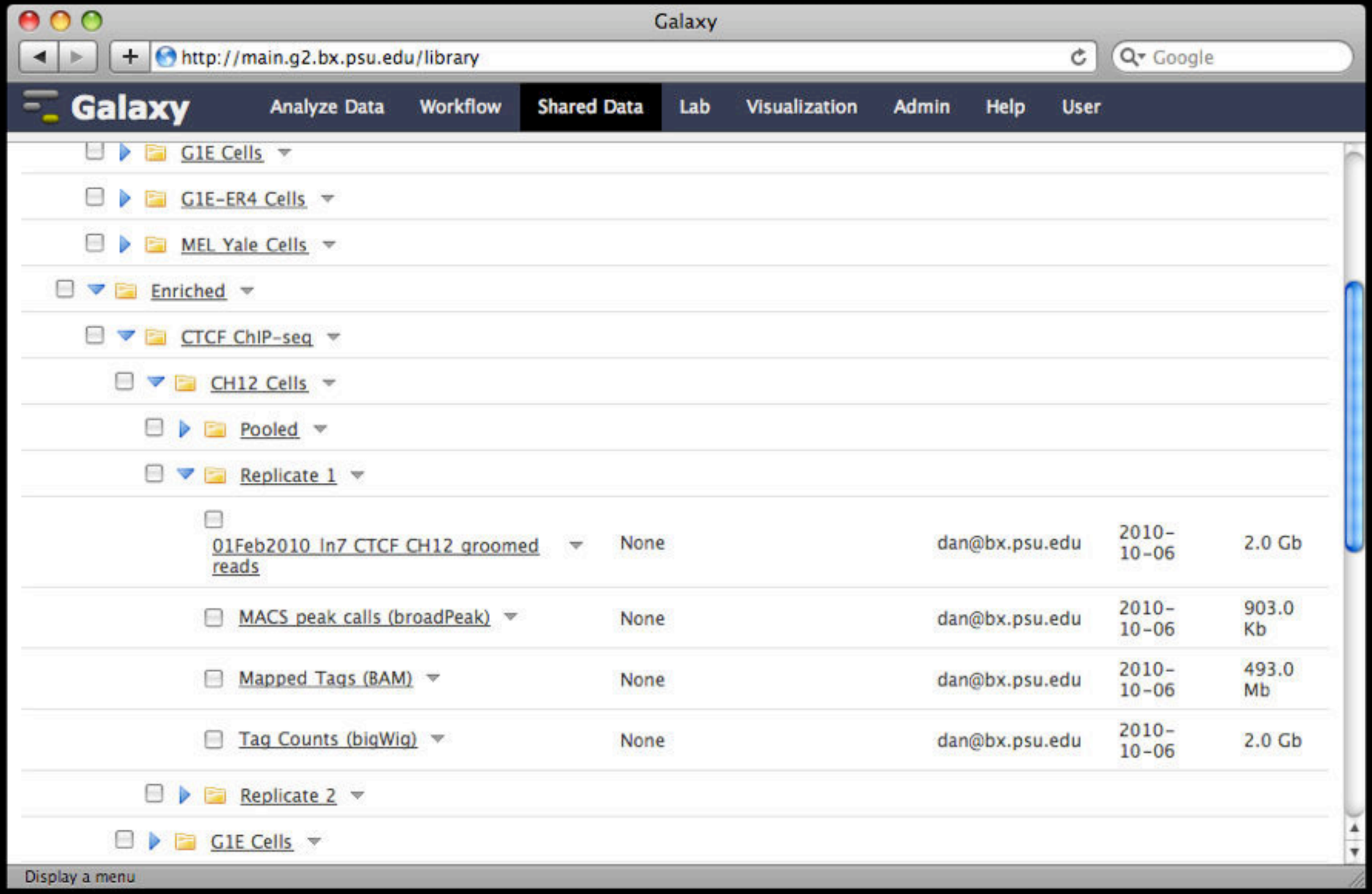
Sharing History with Specific Users

You have not shared this history with any users.

Share with a user

Back to Histories List


Galaxy
http://main.g2.bx.psu.edu/root
Google
Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
Data Libraries
Published Histories
Published Workflows
Published Visualizations
Published Pages
Tools
Options
search tools
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
Human Genome Variation
Advanced fastQ manipulation:
Galactic quickie # 14
454 Mapping: Single End
Galactic quickie # 15
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: $Rev 4802:ea7b055efbfa$
History
Options
Unnamed history
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
Display a menu


Galaxy
http://main.g2.bx.psu.edu/library
Google
Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
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
Display a menu

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)

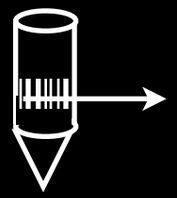
Sample generated by
Cheryl Keller
Who prepared the library? (Optional)

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

Sample Tracking



PACIFIC BIOSCIENCES

illumina

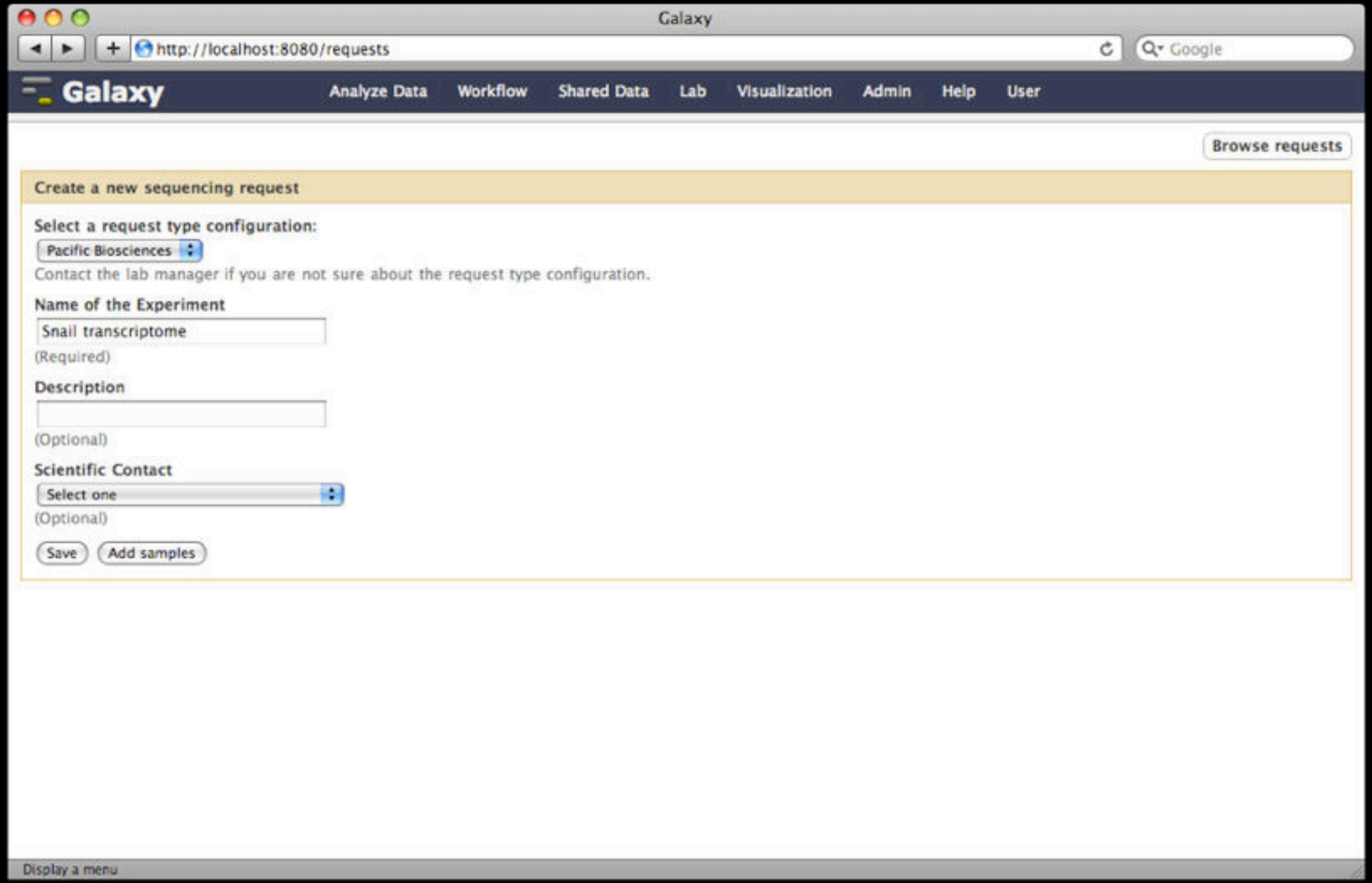
Galaxy Analyze Data

Data Library "Windshield splatter"

Name
Trip A Left Side QV
Trip A Left Side Reads
Trip A Right Side QV
Trip A Right Side Reads
Trip B Left Side QV
Trip B Left Side Reads
Trip B Right Side QV
Trip B Right Side Reads

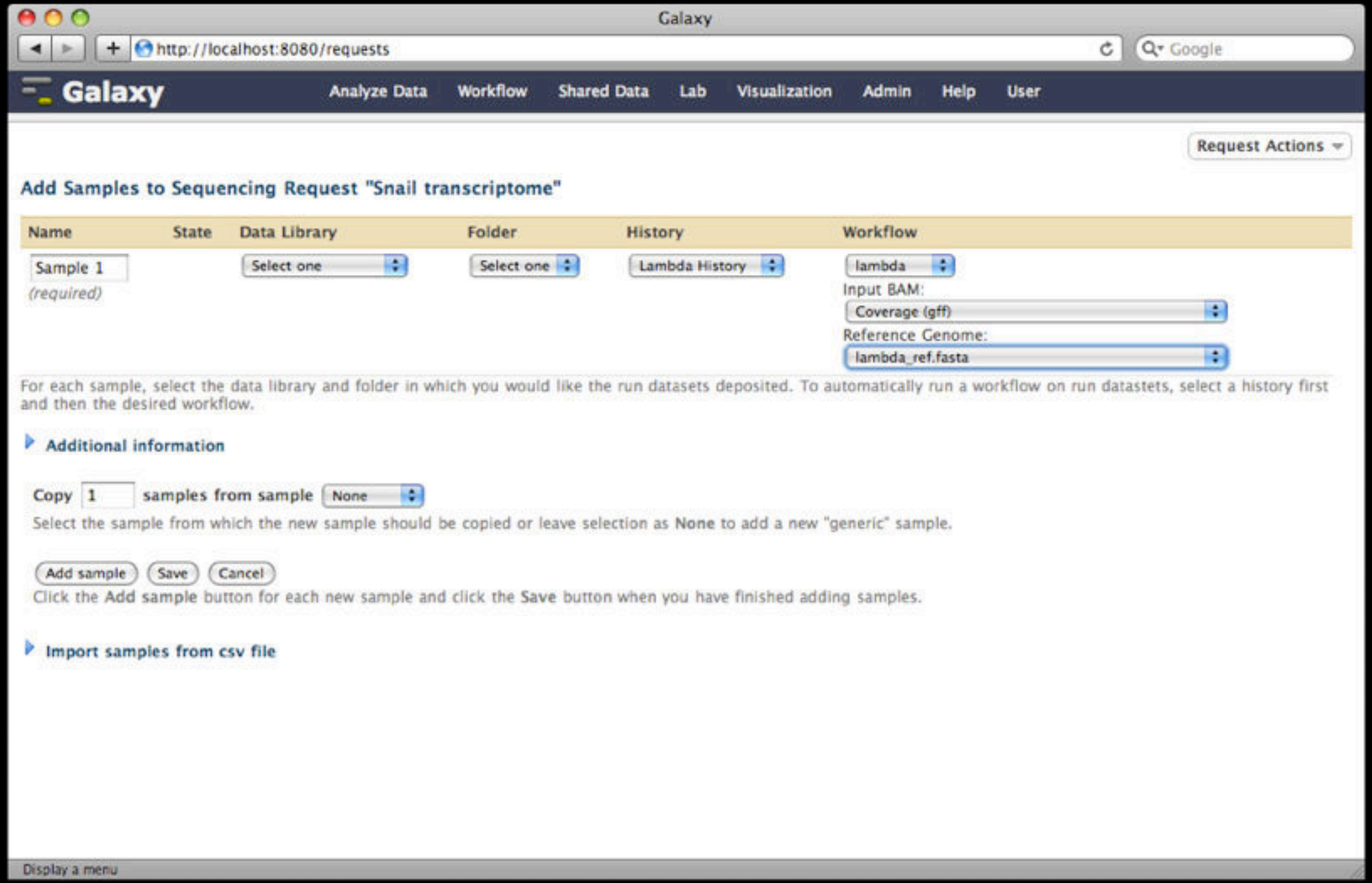
For selected items: Import into your current histo

Lab: Sequencing Request Tracking



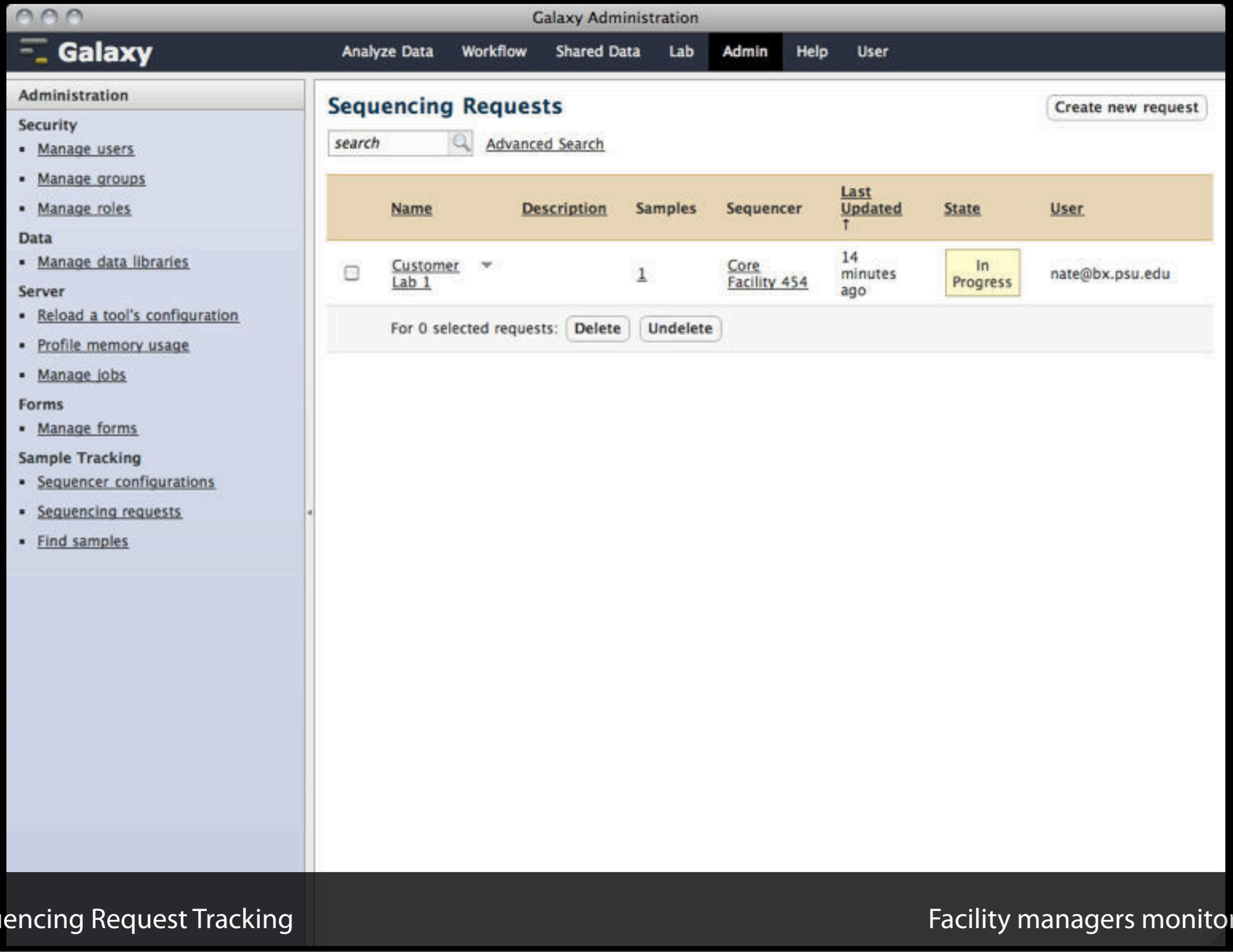
Lab: Sequencing Request Tracking

Customers create their own requests



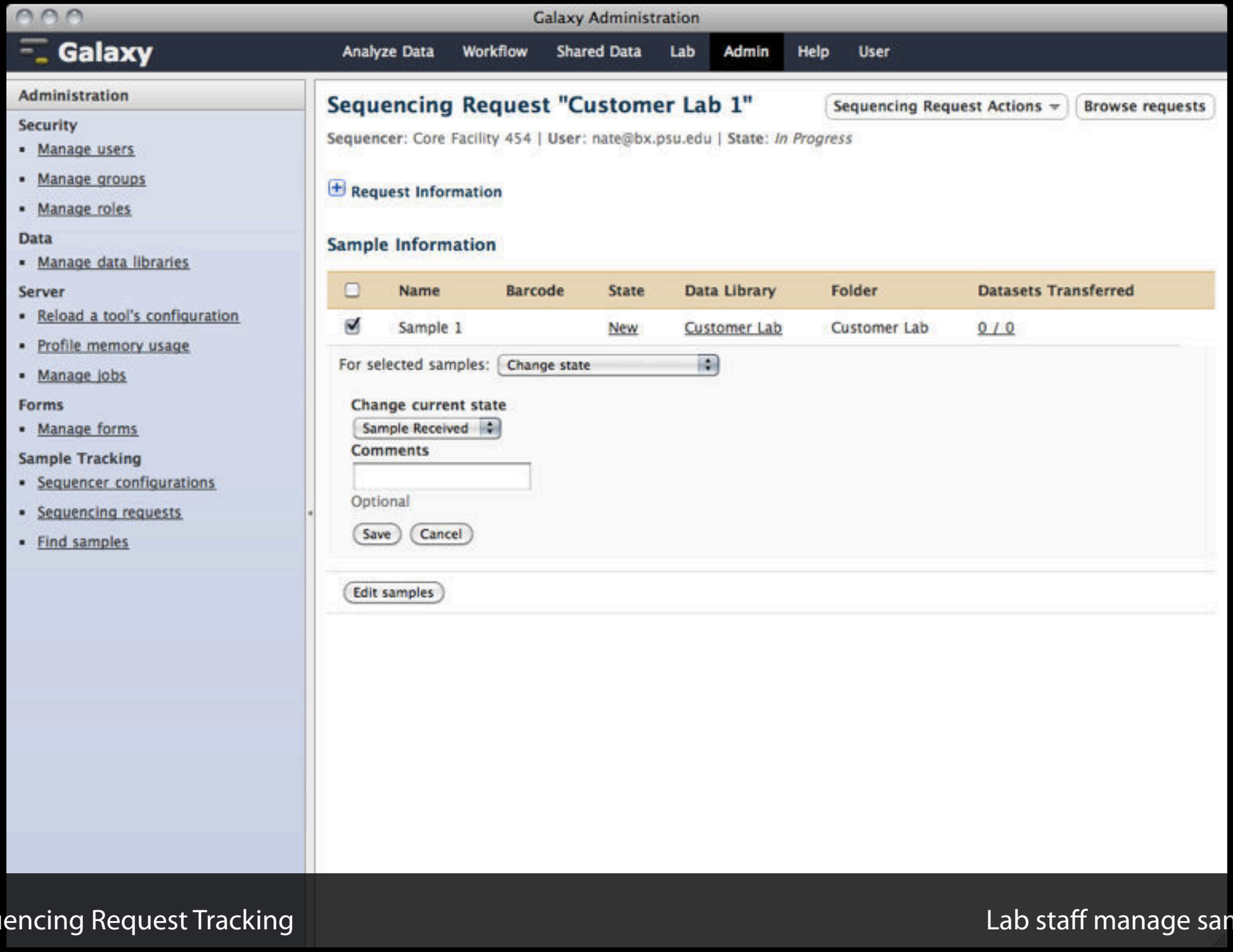
Lab: Sequencing Request Tracking

Customers add samples, cand define library to deposit in and/or workflows to run



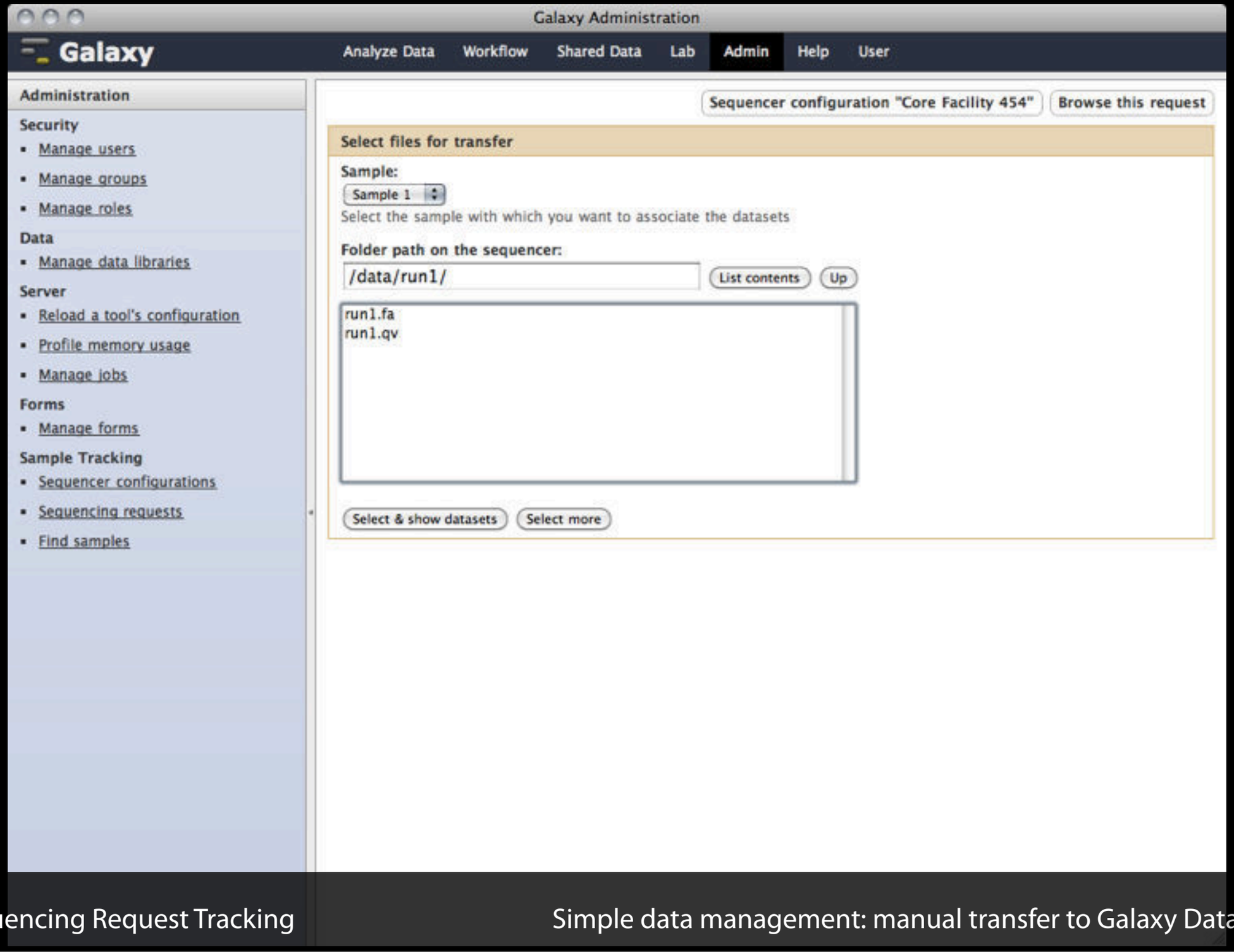
Lab: Sequencing Request Tracking

Facility managers monitor requests



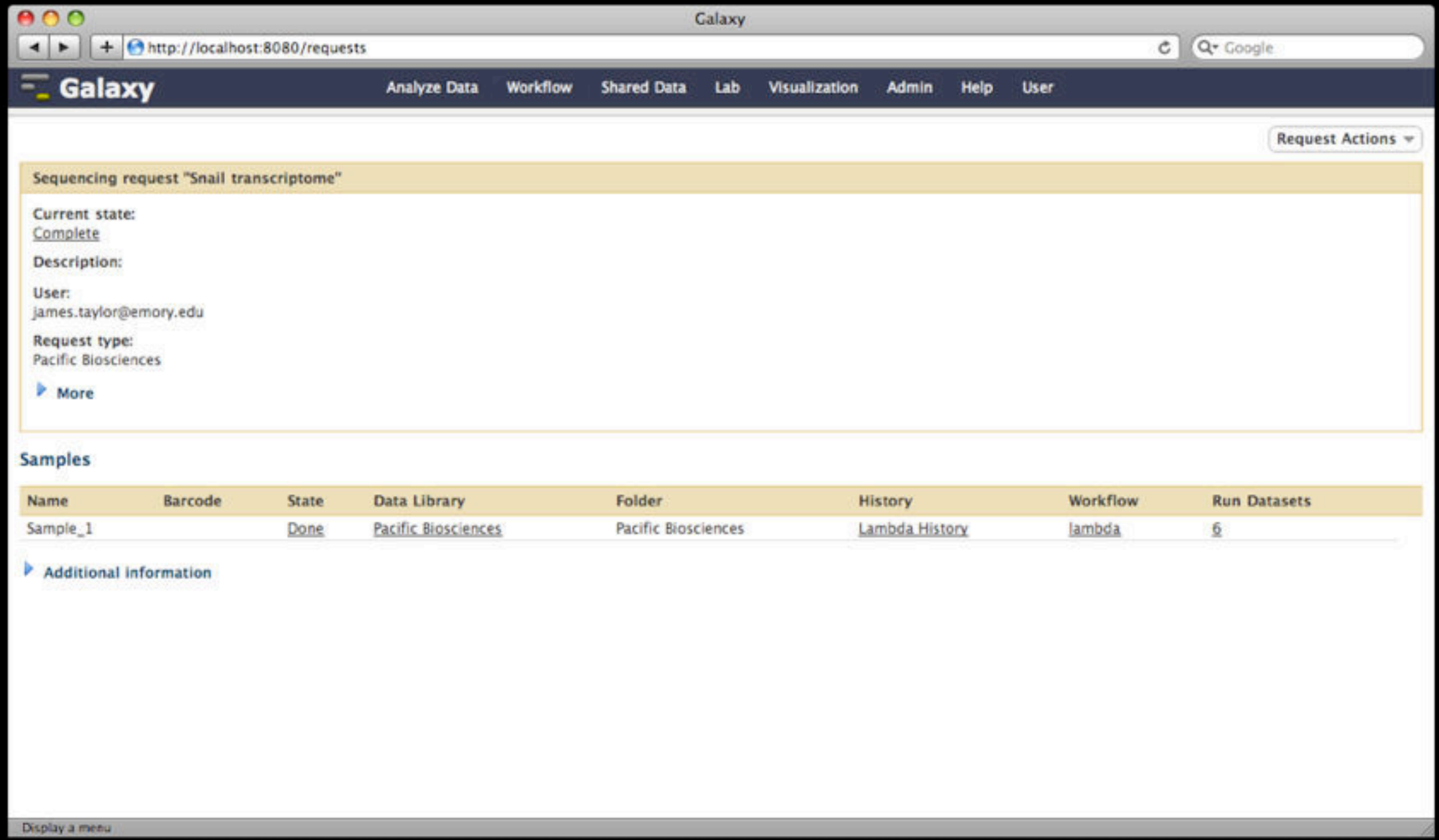
Lab: Sequencing Request Tracking

Lab staff manage sample state



Lab: Sequencing Request Tracking

Simple data management: manual transfer to Galaxy Data Libraries



Lab: Sequencing Request Tracking

Automatic transfer, data loaded into Galaxy and workflow run with no intervention

Sample tracking is completely extensible

- Track manually, with barcodes, or integrate with an existing LIMS
- Everything is configuration driven, capture whatever data and support whatever workflow you want
- Interaction with sequence instruments and secondary analysis is completely pluggable
  - For services that provide a web / REST API even easier

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

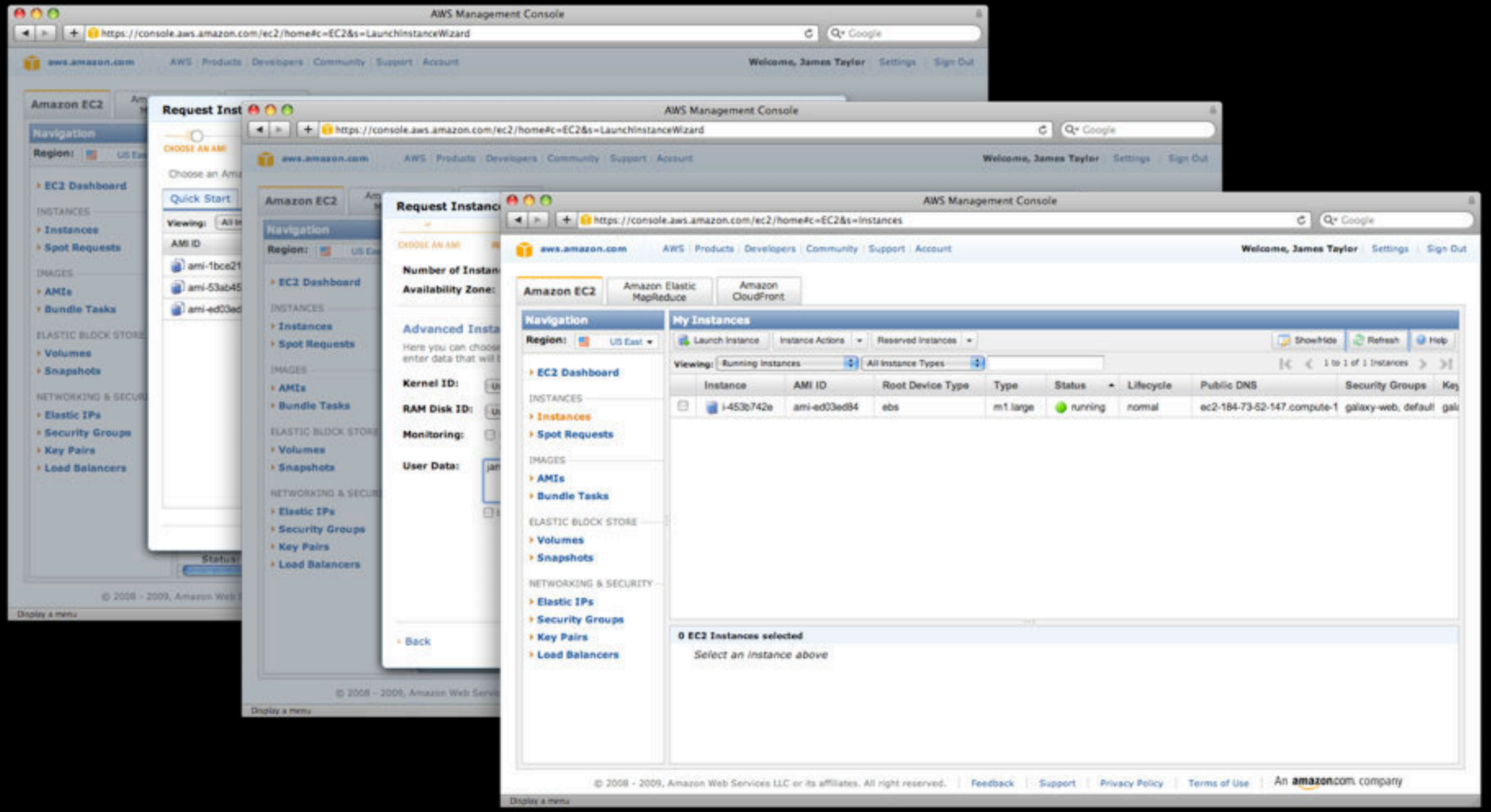
Scale up on your cluster

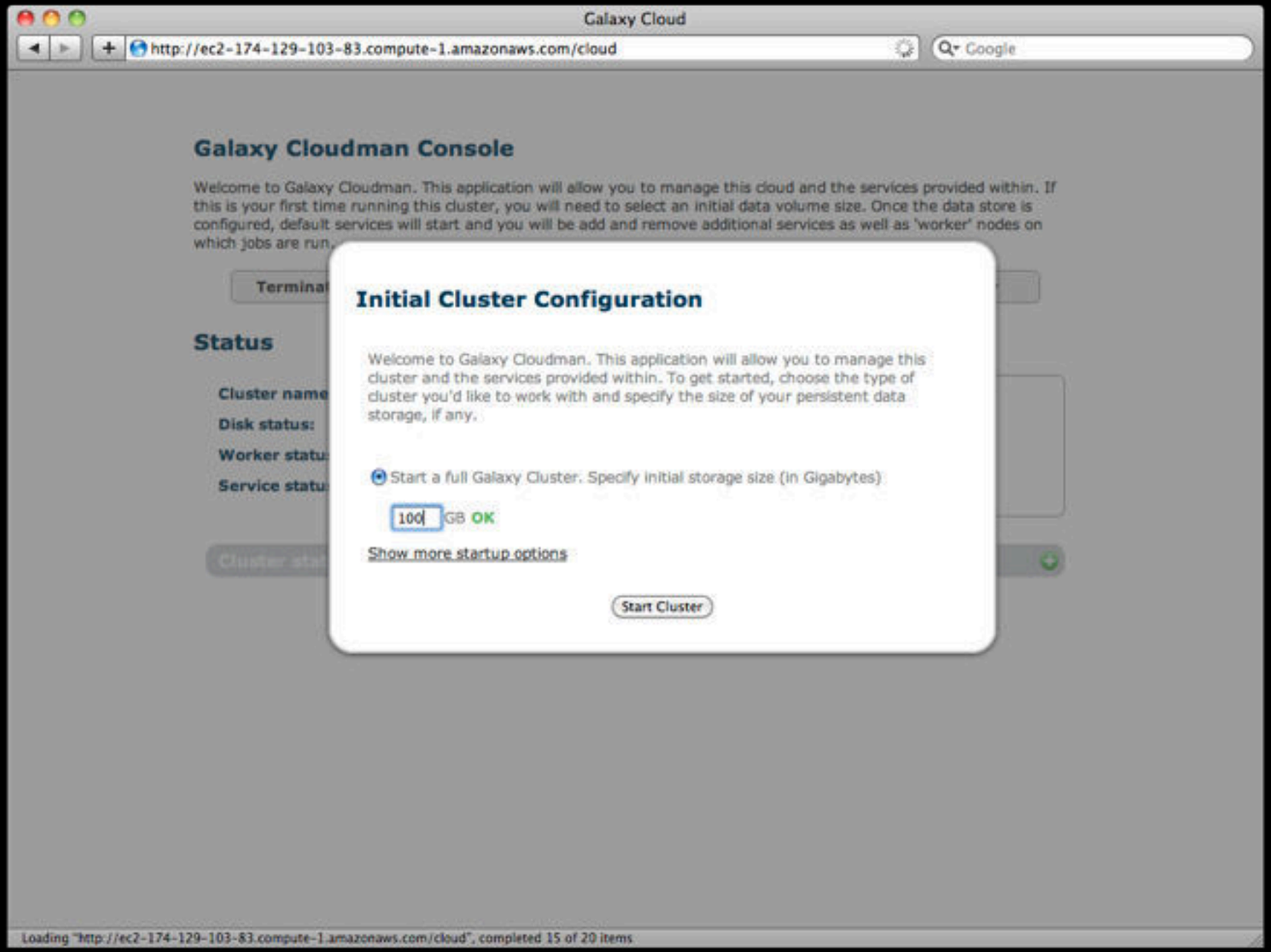
- Move intensive processing (tool execution) to other hosts
- Frees up the application server to serve requests and manage jobs
- Utilize existing resources
- Supports any scheduler that supports DRMAA (most of them)
- It's easy
- But, requires an **existing computational resource** on which to be deployed

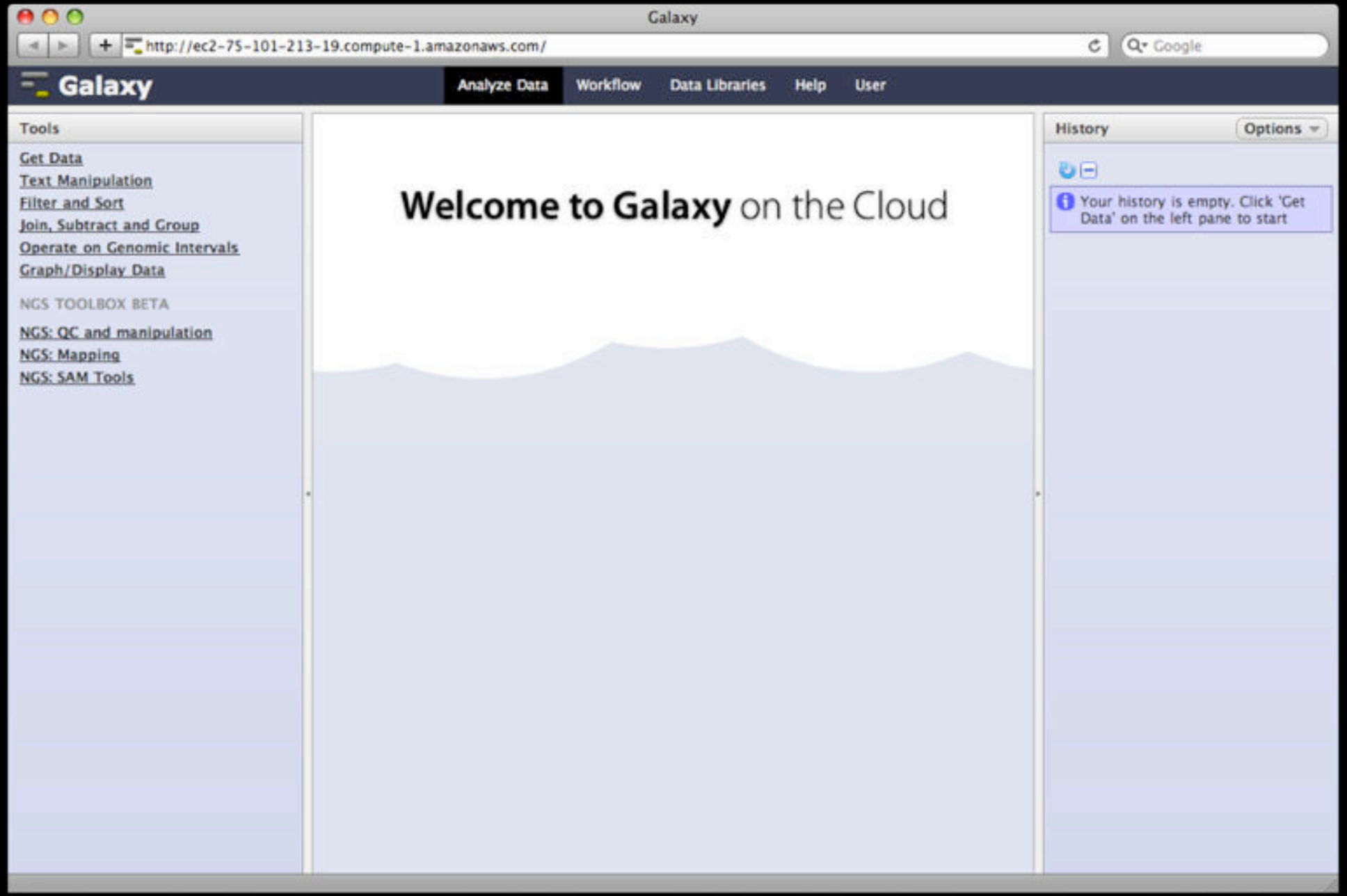
Cloud computing / Infrastructure virtualization

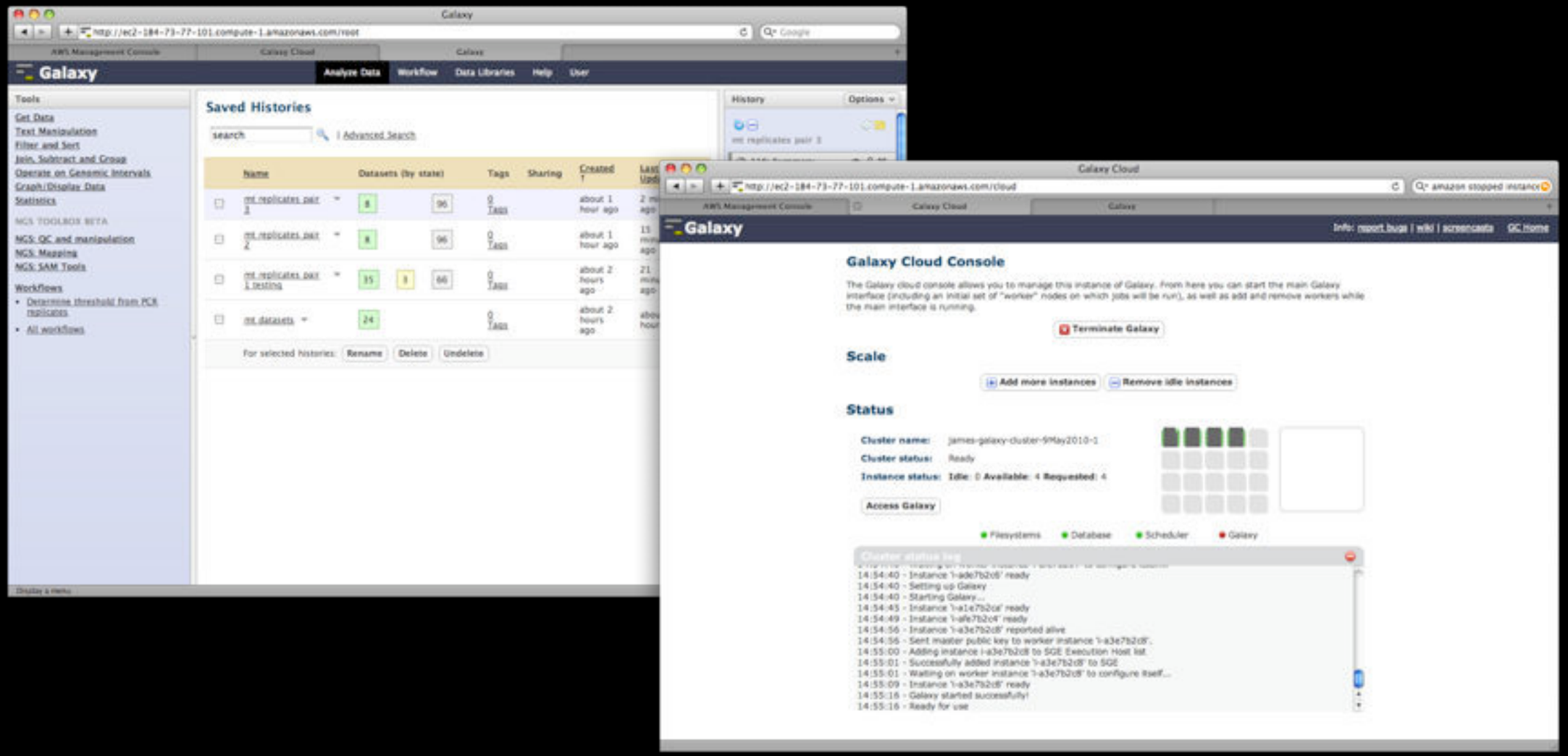
- Computing using resources acquired on demand
- Virtual infrastructure allows for (potential) economies of scale, and (definite) improvements to management automation
- Cloud-style deployment provides a solution both for users without dedicated compute resources, and for simplifying deployment and management

Using Amazon EC2: Startup in 3 steps



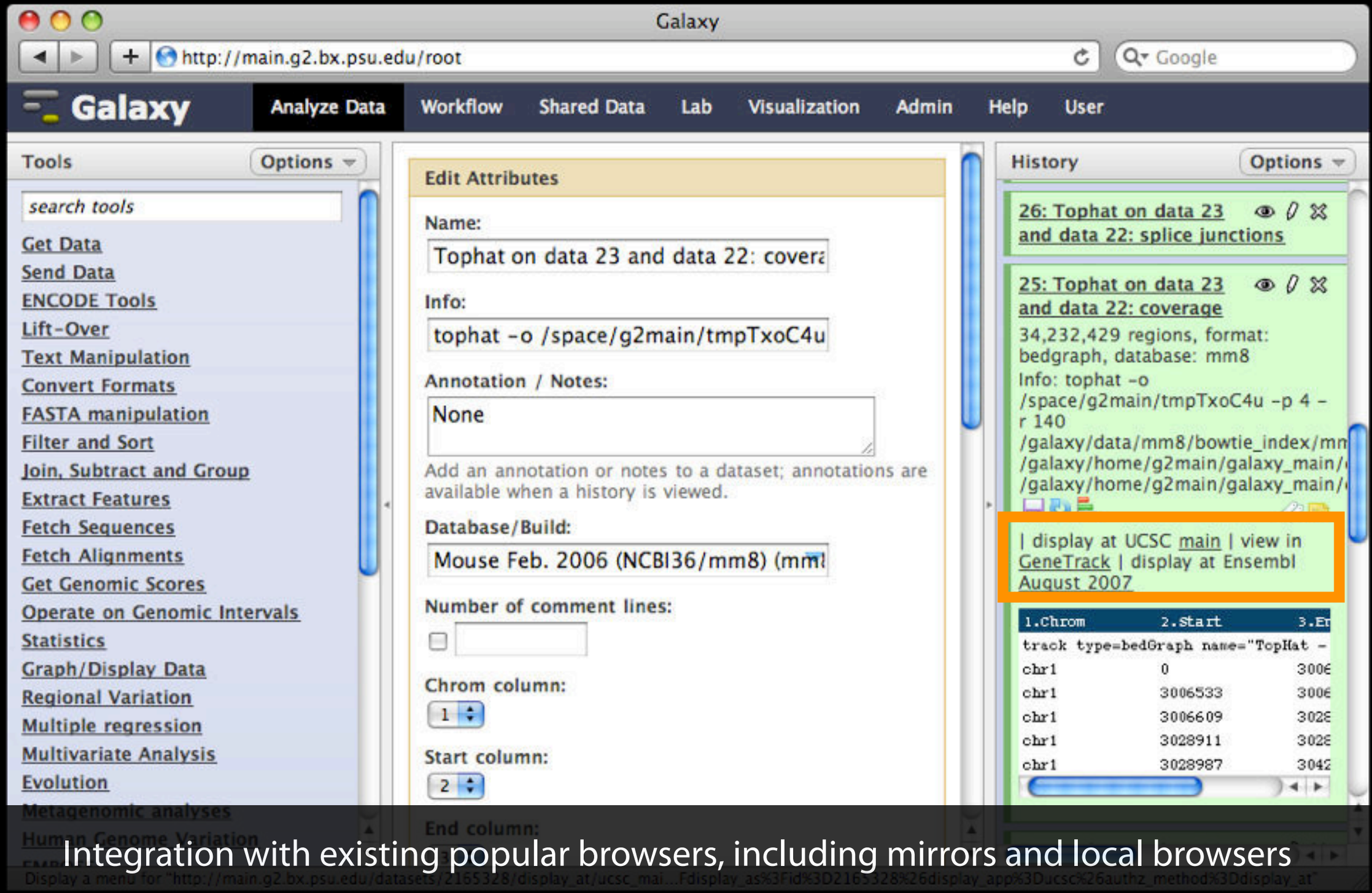

Galaxy Cloud
http://ec2-174-129-103-83.compute-1.amazonaws.com/cloud
Google
Galaxy Cloudman Console
Welcome to Galaxy Cloudman. This application will allow you to manage this cloud and the services provided within. If this is your first time running this cluster, you will need to select an initial data 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.
Terminat
Status
Cluster name
Disk status:
Worker statu
Service statu
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
Show more startup options
Start Cluster
Loading "http://ec2-174-129-103-83.compute-1.amazonaws.com/cloud", completed 15 of 20 items


Galaxy
http://ec2-75-101-213-19.compute-1.amazonaws.com/
Google
Galaxy
Analyze Data
Workflow
Data Libraries
Help
User
Tools
Get Data
Text Manipulation
Filter and Sort
Join, Subtract and Group
Operate on Genomic Intervals
Graph/Display Data
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
Welcome to Galaxy on the Cloud
History
Options
Your history is empty. Click 'Get Data' on the left pane to start

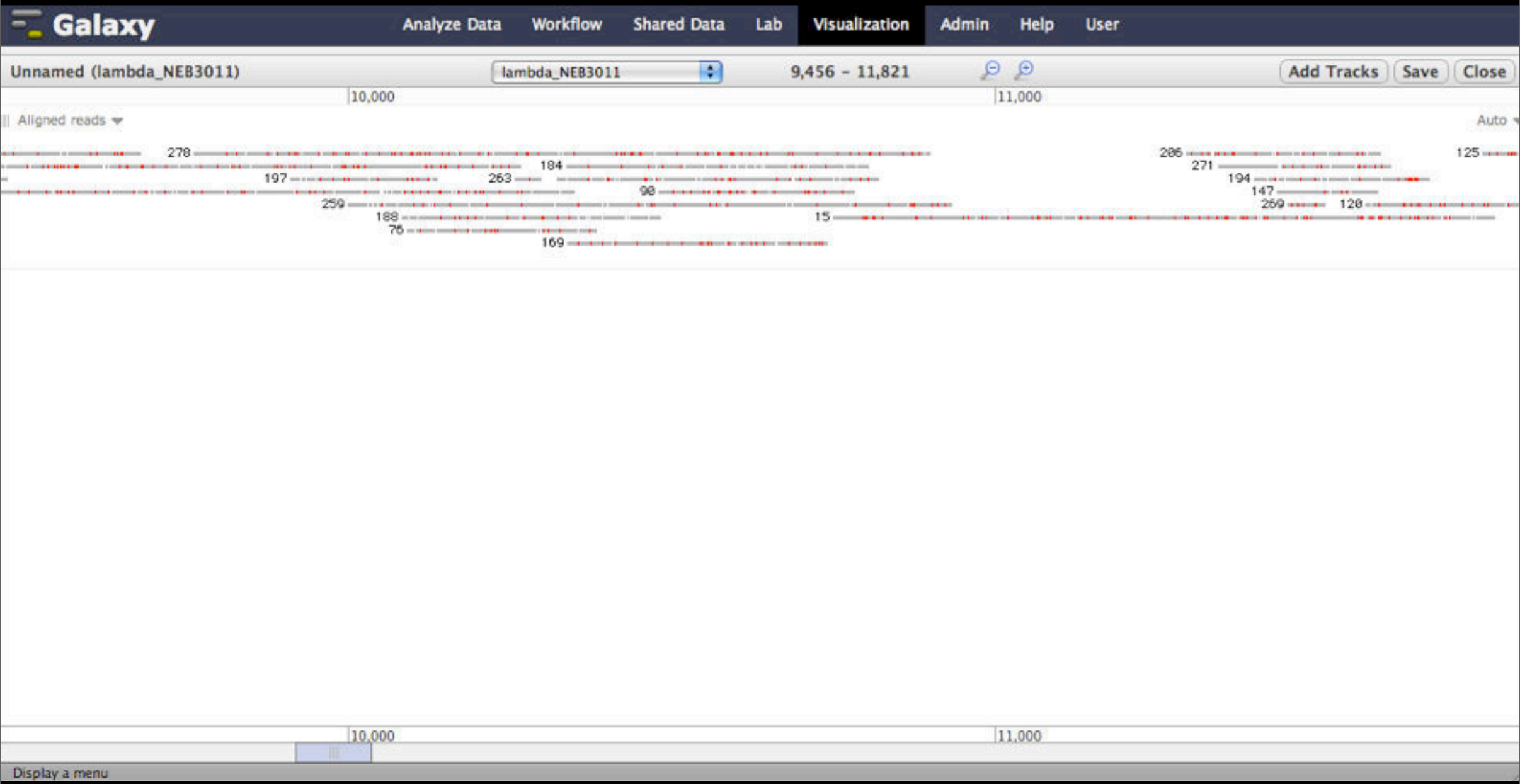


Can use like any other Galaxy instance, with additional compute nodes acquired and released (*automatically*) in response to usage

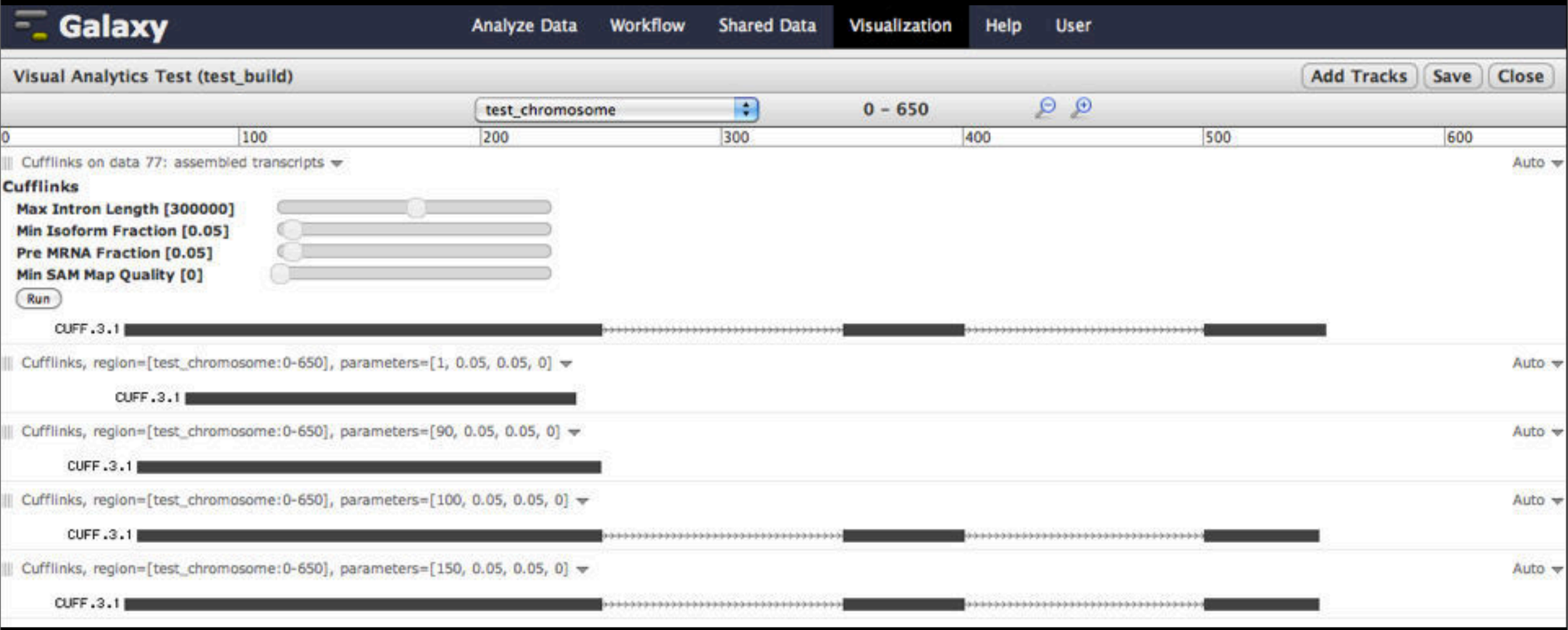
Visualization
(beta)



Integration with existing popular browsers, including mirrors and local browsers



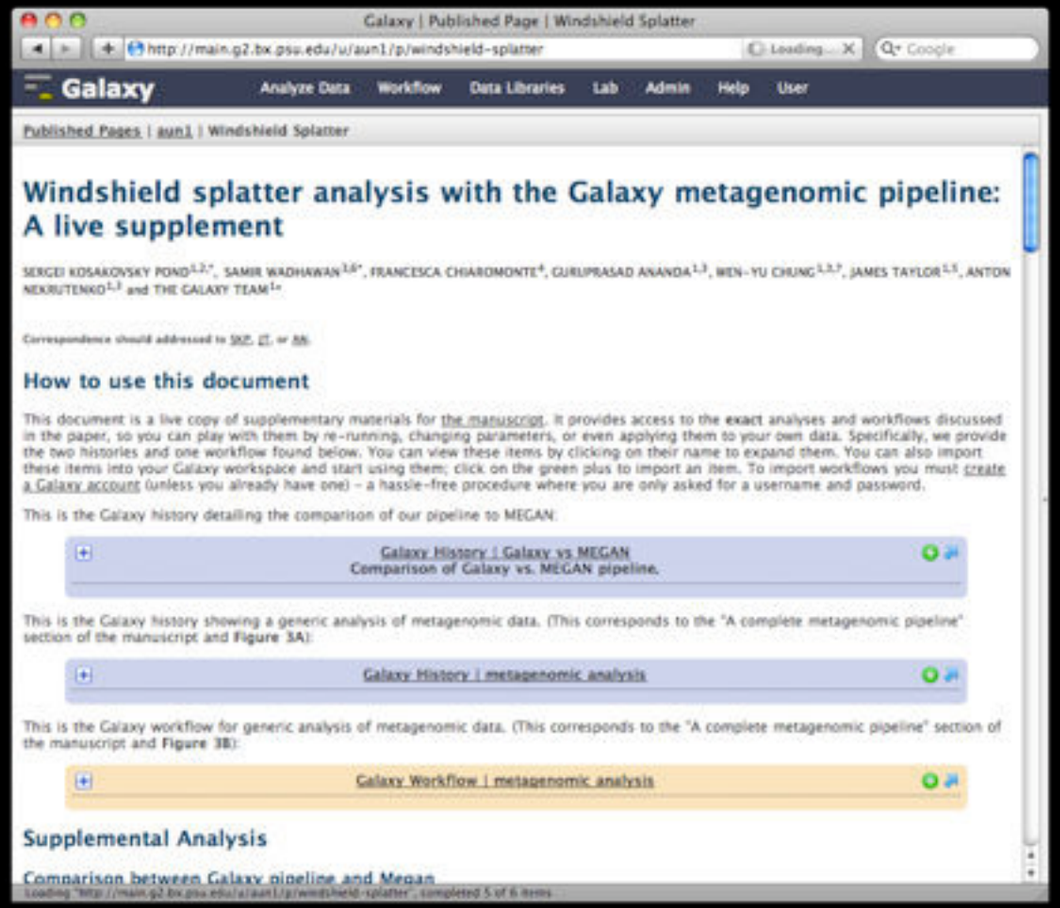
Visualizing aligned reads in trackster



Visualization integrated with tools: visual analytics in trackster

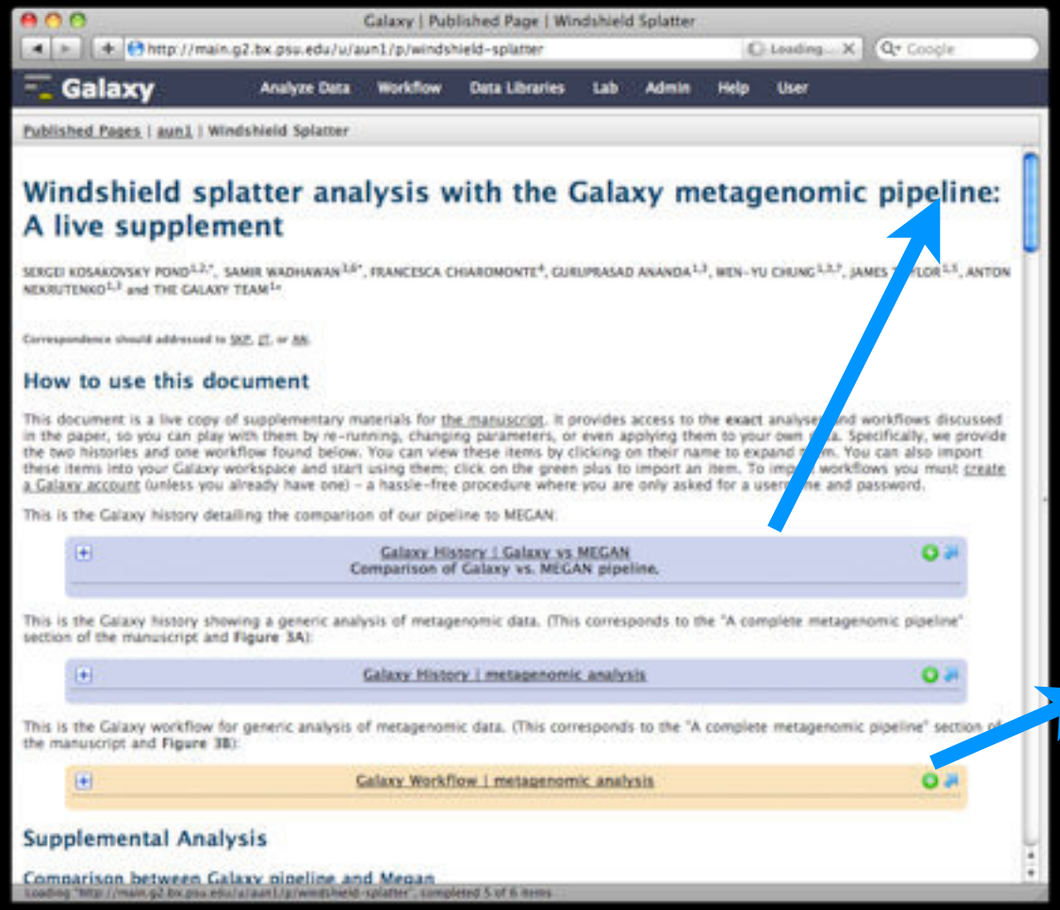
Publishing analysis

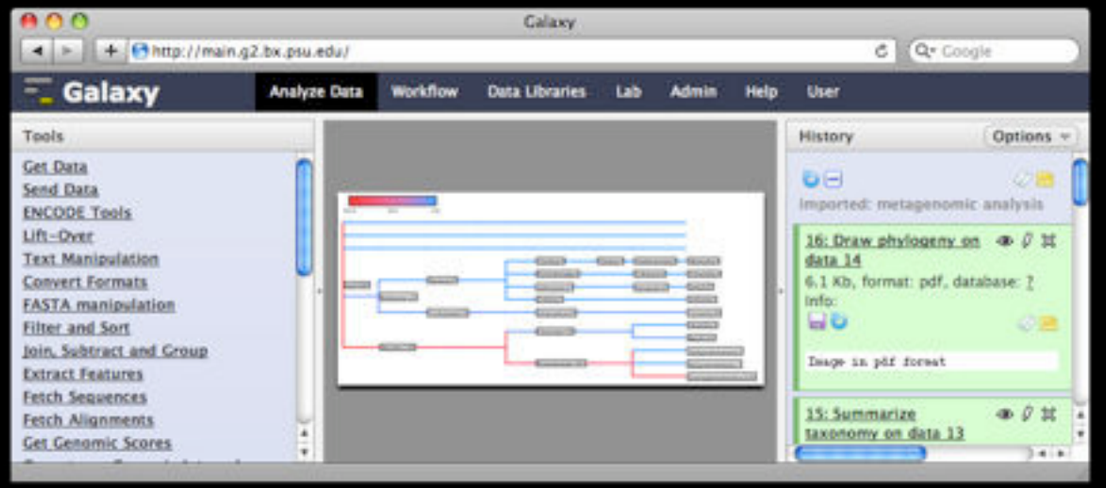
Sharing and publishing

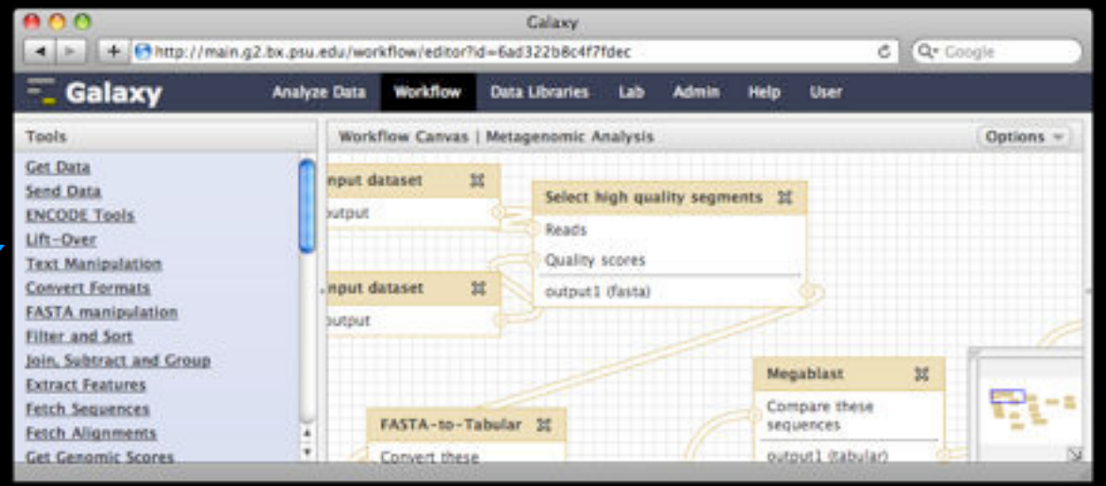


- All analysis components (datasets, histories, workflows) can be *shared* among Galaxy users and *published*
- Pages and annotation allow analysis to be augmented with textual content and provided in the form of an integrated document

Sharing and publishing







Try it now:

http://usegalaxy.org

Develop and deploy:

http://getgalaxy.org