Utilizing the Galaxy Analysis Framework at Core Facilities

Western Association of Core Directors (WACD)
September 18, 2015

Dave Clements
Galaxy Team
Johns Hopkins University
http://galaxyproject.org/

#usegalaxy @galaxyproject

Talk plan when I got off the train yesterday morning

1/3 What is Galaxy and what can it do?
1/3 Help clients to do their own data analysis w/ Galaxy!
1/6 Using Galaxy for in-house data pipelines
1/6 Q & A

http://galaxyproject.org

Talk plan after yesterday

~~**1/3** What is Galaxy and what can it do?~~
~~**1/3** Help clients to do their own data analysis w/ Galaxy!~~
~~**1/6** Using Galaxy for in-house data pipelines~~
~~**1/6** Q & A~~

2/10 What is Galaxy and what can it do?
1/10 Using Galaxy for in-house data pipelines
2/10 Help clients to do their own data analysis w/ Galaxy!
5/10 Open discussion:

What is the role of cores in supporting client data analysis?
Should this be part of your value proposition?

http://galaxyproject.org

What is Galaxy?

Data integration and analysis platform that emphasizes accessibility, reproducibility, and transparency

http://galaxyproject.org

What is Galaxy?

Keith Bradnam's definition:

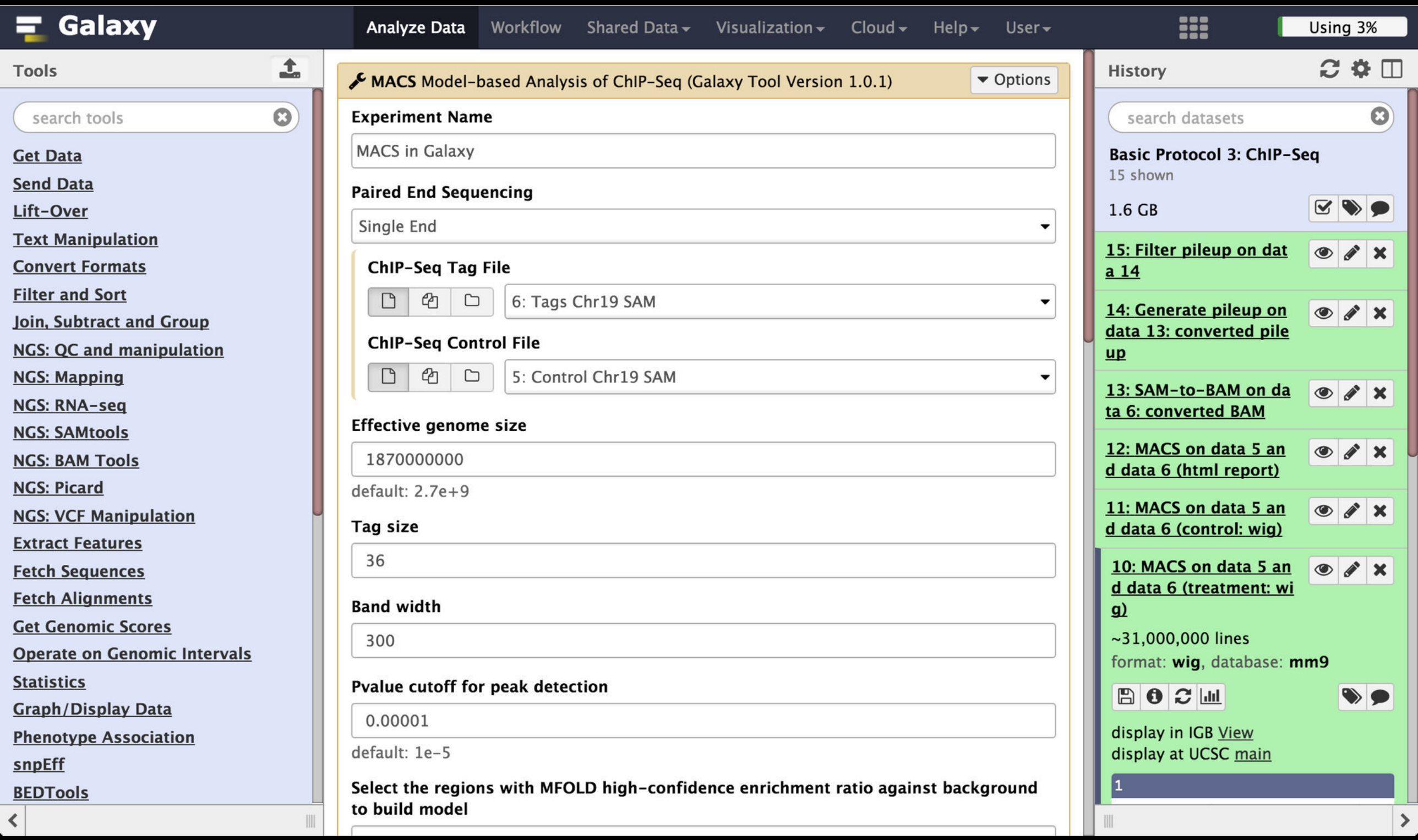
"A web-based platform that provides a simplified interface to many popular bioinformormatics tools."

From

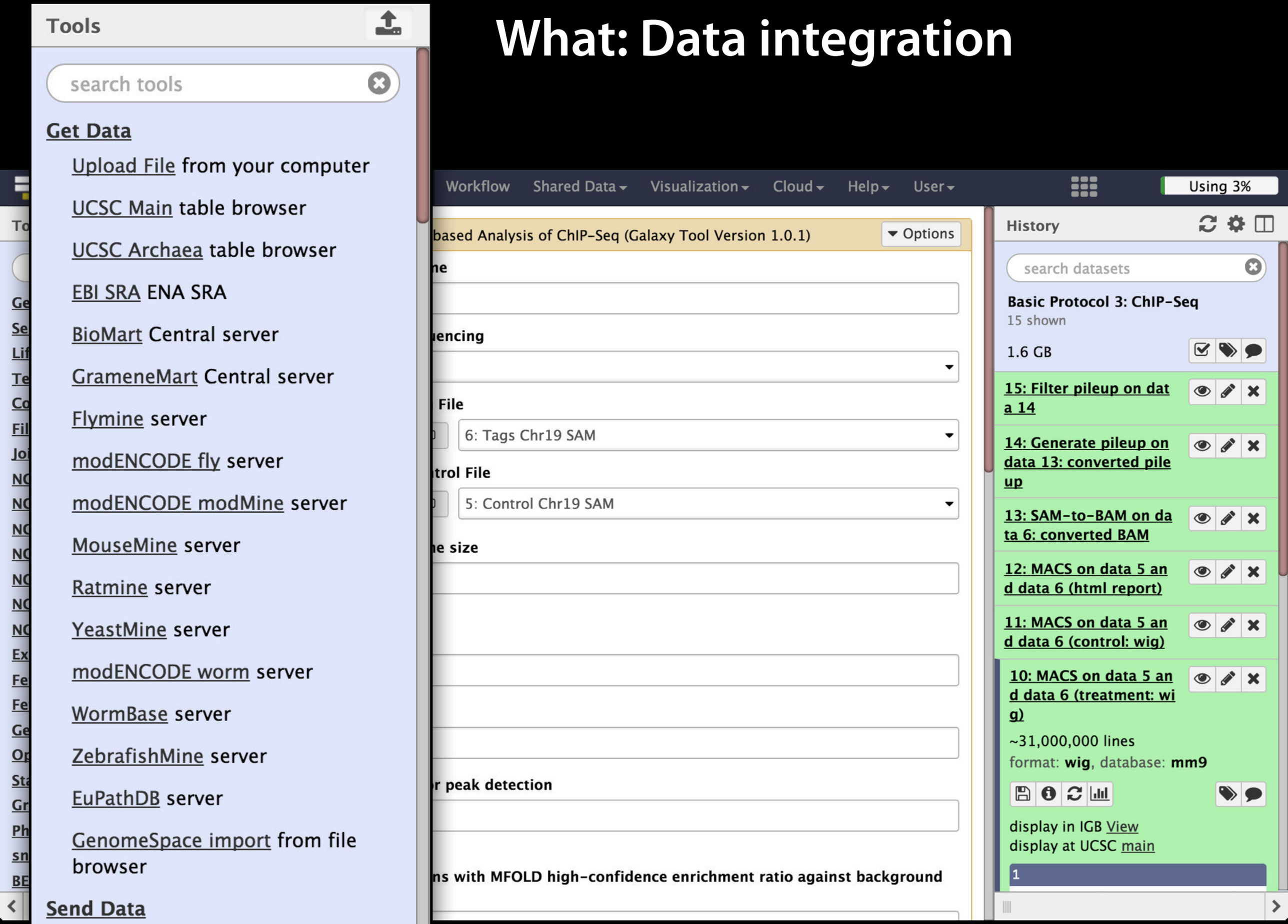
"13 Questions You May Have About Galaxy"

http://bit.ly/13questions

What: A web based platform



What: Data integration



What: Framework for Tools

NGS: QC and manipulation

- FastQC Read Quality reports
- Select high quality segments
- Build base quality distribution
- Draw quality score boxplot
- Quality format converter (ASCII-Numeric)
- Filter by quality
- FASTQ to FASTA converter
- Remove sequencing artifacts
- Barcode Splitter
- Clip adapter sequences
- Collapse sequences
- Draw nucleotides distribution chart
- Compute quality statistics
- Rename sequences
- Reverse-Complement
- Trim sequences
- Combine FASTA and QUAL into FASTQ
- Filter FASTQ reads by quality score and length
- Manipulate FASTQ reads on various attributes
- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ Masker by quality score
- FASTQ joiner on paired end reads
- FASTQ splitter on joined paired end reads
- FASTQ Summary Statistics by column
- FASTQ to FASTA converter
- FASTQ to Tabular converter
- FASTQ Trimmer by column
- FASTQ Quality Trimmer by sliding window
- Tabular to FASTQ converter
- Convert SOLiD output to fastq
- Compute quality statistics for SOLiD data
- Draw quality score boxplot for SOLiD data

NGS: Mapping

- Bowtie2 - map reads against reference genome

loud Help User Using 3%

0.1) Options

o against background

History

search datasets

Basic Protocol 3: ChIP-Seq
15 shown

1.6 GB

15: Filter pileup on dat a 14

14: Generate pileup on data 13: converted pile up

13: SAM-to-BAM on da ta 6: converted BAM

12: MACS on data 5 an d data 6 (html report)

11: MACS on data 5 an d data 6 (control: wig)

10: MACS on data 5 an d data 6 (treatment: wi g)

~31,000,000 lines
format: **wig**, database: **mm9**

display in IGB View
display at UCSC main

1

https://toolshed.g2.bx.psu.edu

Galaxy Tool Shed

Repositories Groups Help User

3548 valid tools on Aug 31, 2015

Search

- Search for valid tools
- Search for workflows

Valid Galaxy Utilities

- Tools
- Custom datatypes
- Repository dependency definitions
- Tool dependency definitions

All Repositories

- Browse by category

Available Actions

- Login to create a repository

Repositories by Category

search repository name, description

Name	Description	Repositories
Assembly	Tools for working with assemblies	53
ChIP-seq		
Combinatorial Selections		
Computational chemistry		
Convert Formats		
Data Managers		
Data Source		
Fasta Manipulation		
Fastq Manipulation		
Genome-Wide Association Study		
Genomic Interval Operations		
Graphics		
Imaging	Utilities to support imaging	
Metabolomics	Tools for use in the study of Metabolomics	
Metagenomics	Tools enabling the study of metagenomes	
Micro-array Analysis	Tools for performing micro-array analysis	
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencin	
Ontology Manipulation	Tools for manipulating ontologies	
Phylogenetics	Tools for performing phylogenetic analysis	
Proteomics	Tools enabling the study of proteins	
RNA	Utilities for RNA	
SAM	Tools for manipulating alignments in the SAM format	
Sequence Analysis	Tools for performing Protein and DNA/RNA analysis	
Statistics	Tools for generating statistics	
Systems Biology	Systems biology tools	
Text Manipulation	Tools for manipulating data	
Tool Dependency Packages	Repositories that contain third-party tool dependency pack installation definitions	
Tool Generators	Tools that make or help make new tools	

What: Framework for Tools

https://toolshed.g2.bx.psu.edu

Galaxy Tool Shed

Repositories Groups Help User

3548 valid tools on Aug 31, 2015

Search

- Search for valid tools
- Search for workflows

Valid Galaxy Utilities

- Tools
- Custom datatypes
- Repository dependency definitions
- Tool dependency definitions

All Repositories

- Browse by category

Available Actions

- Login to create a repository

Repositories by Category

search repository name, description

Name	Description	Repositories
Assembly	Tools for working with assemblies	53
ChIP-seq		
Combinatorial Selections		
Computational chemistry		
Convert Formats		
Data Managers		
Data Source		
Fasta Manipulation		
Fastq Manipulation		
Genome-Wide Association Study		
Genomic Interval Operations		
Graphics		
Imaging	Utilities to support imaging	
Metabolomics	Tools for use in the study of Metabolomics	
Metagenomics	Tools enabling the study of metagenomes	
Micro-array Analysis	Tools for performing micro-array analysis	
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencin	
Ontology Manipulation	Tools for manipulating ontologies	
Phylogenetics	Tools for performing phylogenetic analysis	
Proteomics	Tools enabling the study of proteins	
RNA	Utilities for RNA	
SAM	Tools for manipulating alignments in the SAM format	
Sequence Analysis	Tools for performing Protein and DNA/RNA analysis	
Statistics	Tools for generating statistics	
Systems Biology	Systems biology tools	
Text Manipulation	Tools for manipulating data	
Tool Dependency Packages	Repositories that contain third-party tool dependency pack installation definitions	
Tool Generators	Tools that make or help make new tools	

What: Framework for Tools

Have mass spec data?

Galaxy-P!

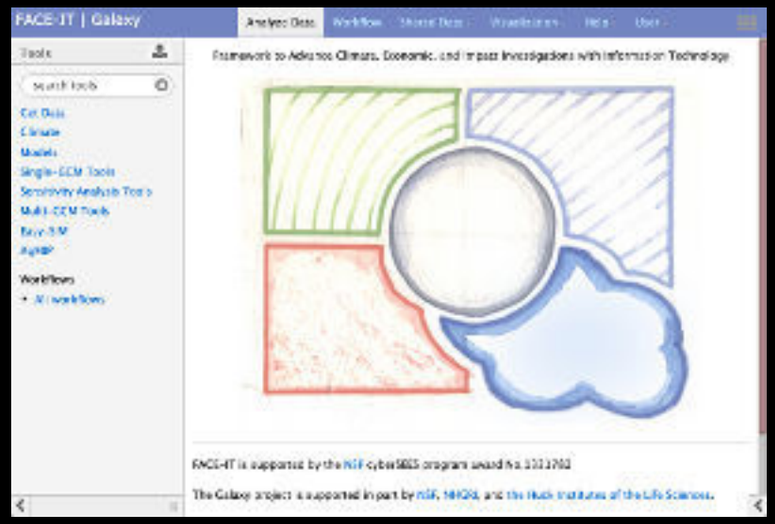
Three ways to Galaxy-P...

Public usegalaxyp.org

Local getgalaxyp.org

Cloud biocloudcentral.msi.umn.edu

or maybe 4... Twitter @usegalaxyp



OPEN SOURCE DRUG DISCOVERY

OSDDLinux LiveGalaxy

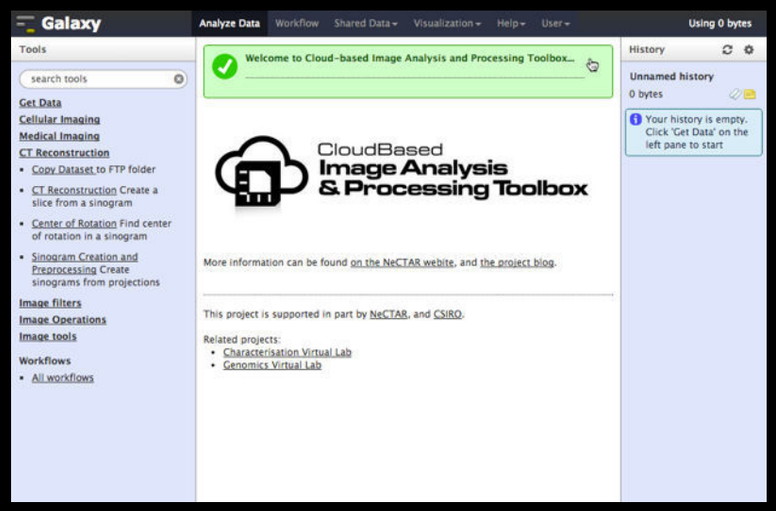
ballaxy

Powered by the Biochemical Algorithms Library Project

SYMMETRY

Got symmetry?

Find out.



Proteomics

Metabolomics

Drug Discovery

Cosmology

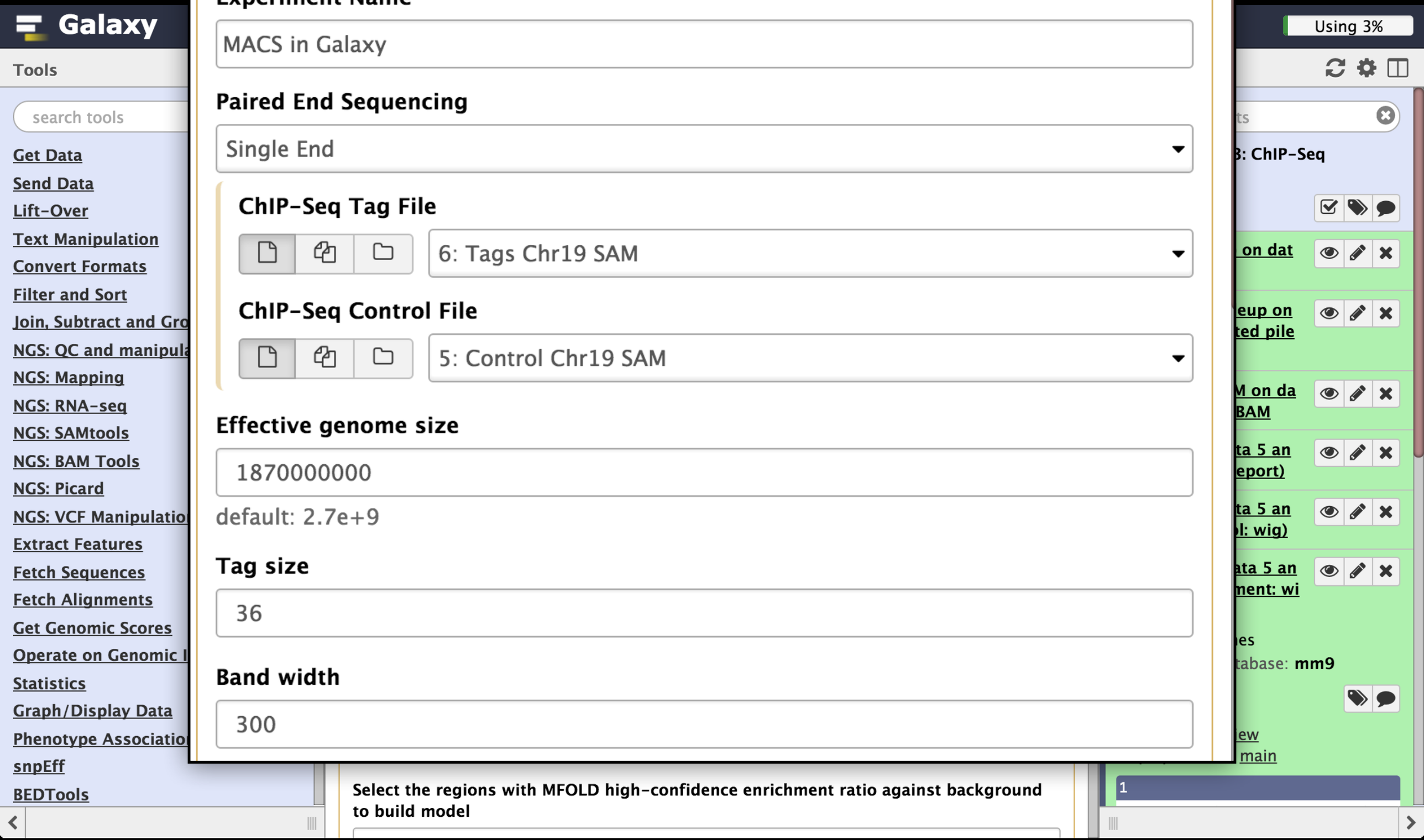
Image Analysis

Social Science

Climate Change

Natural Language

What: Uniform interface to Tools



What: Automatically records all aspects of analysis

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud

Tools

search tools

- Get Data
- Send Data
- Lift-Over
- Text Manipulation
- Convert Formats
- Filter and Sort
- Join, Subtract and Group
- NGS: QC and manipulation
- NGS: Mapping
- NGS: RNA-seq
- NGS: SAMtools
- NGS: BAM Tools
- NGS: Picard
- NGS: VCF Manipulation
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Phenotype Association
- snpEff
- BEDTools

MACS Model-based Analysis of ChIP-Seq (Galaxy Tool Version 1.0.1)

Experiment Name

MACS in Galaxy

Paired End Sequencing

Single End

ChIP-Seq Tag File

6: Tags Chr19 SAM

ChIP-Seq Control File

5: Control Chr19 SAM

Effective genome size

1870000000

default: 2.7e+9

Tag size

36

Band width

300

Pvalue cutoff for peak detection

0.00001

default: 1e-5

Select the regions with MFOLD high-confidence enrichment ratio against to build model

History

search datasets

Basic Protocol 3: ChIP-Seq

15 shown

1.6 GB

15: Filter pileup on data 14

14: Generate pileup on data 13: converted pile up

13: SAM-to-BAM on data 6: converted BAM

12: MACS on data 5 and data 6 (html report)

11: MACS on data 5 and data 6 (control: wig)

10: MACS on data 5 and data 6 (treatment: wig)

~31,000,000 lines

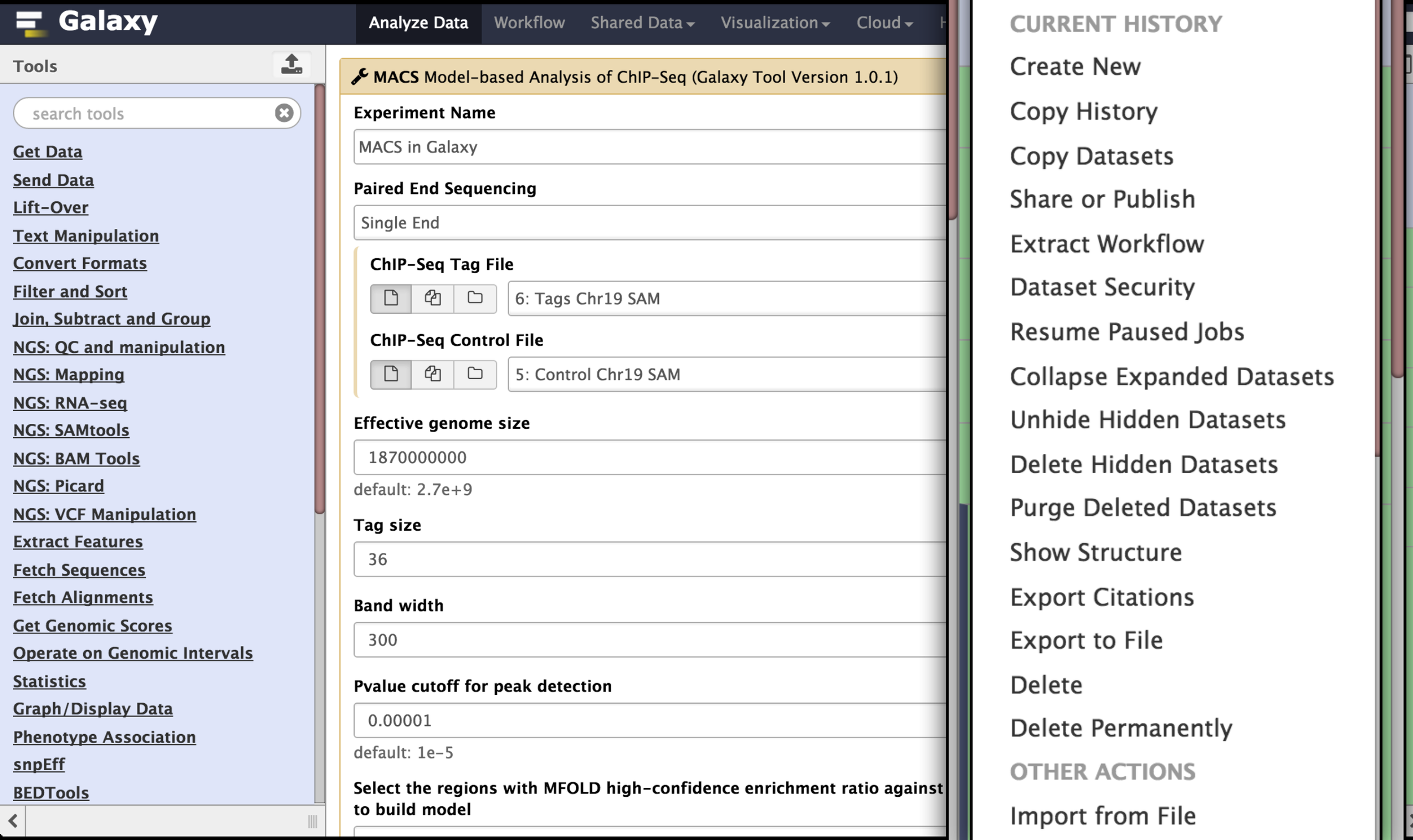
format: **wig**, database: **mm9**

display in IGB View

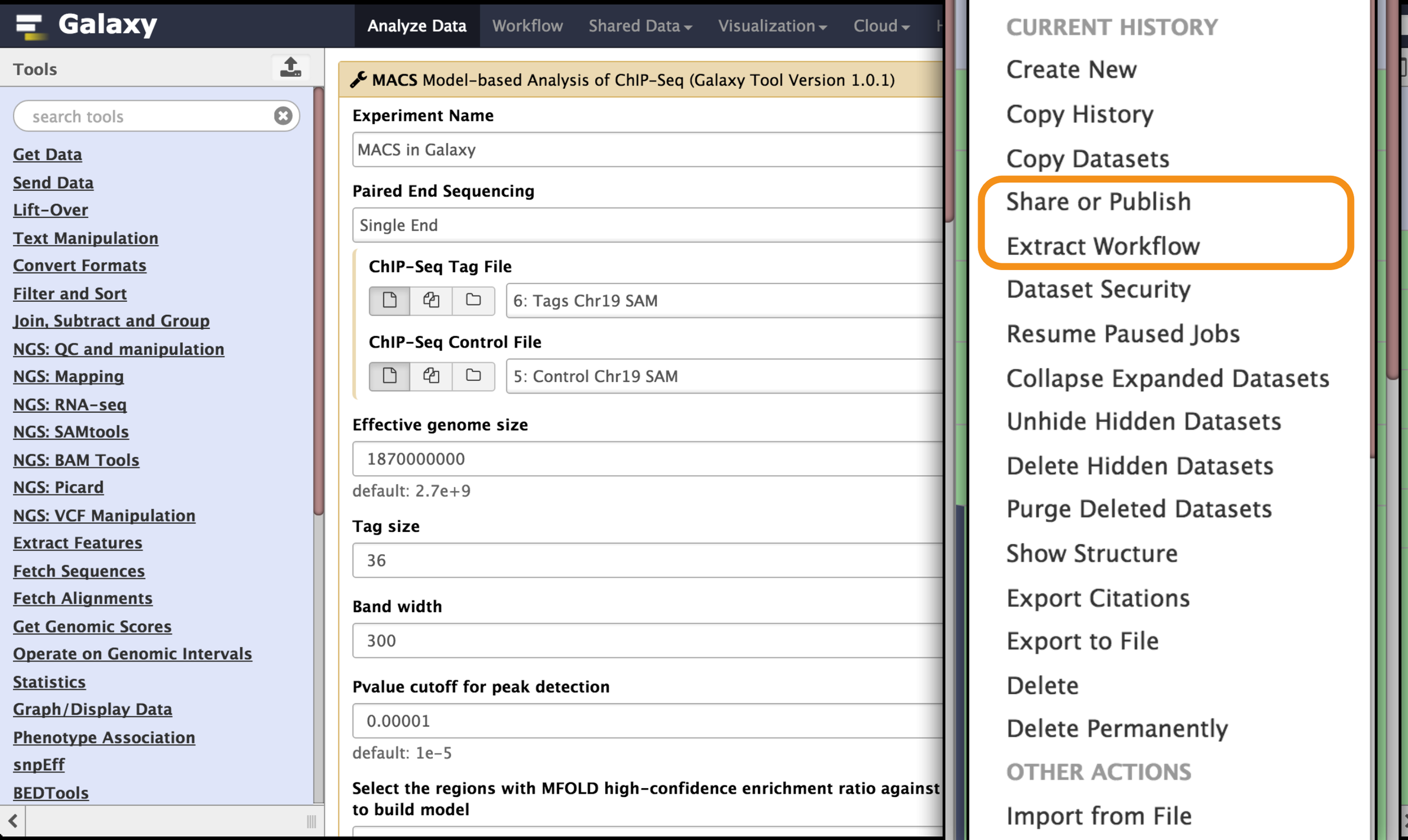
display at UCSC main

1

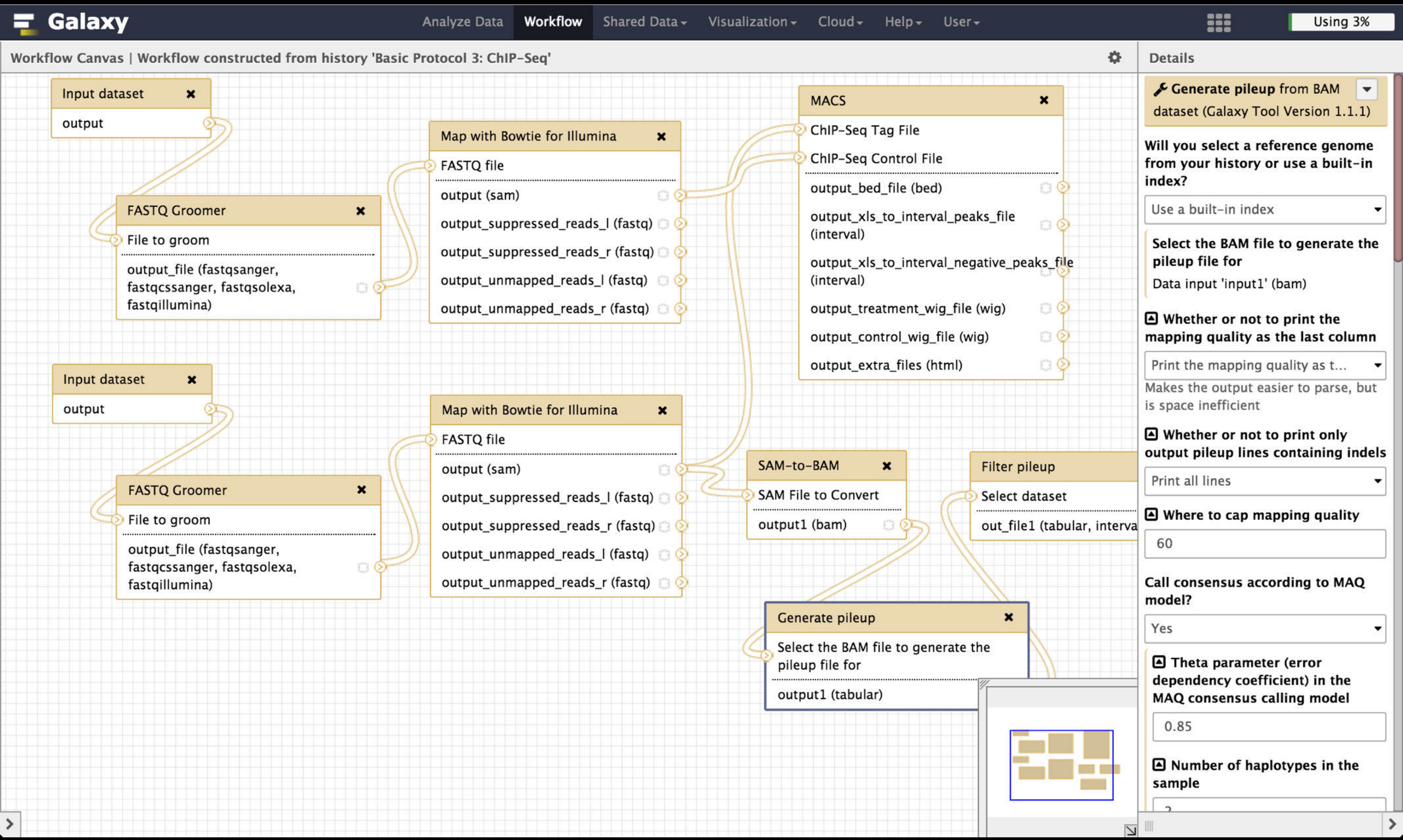
What: History / Analysis management



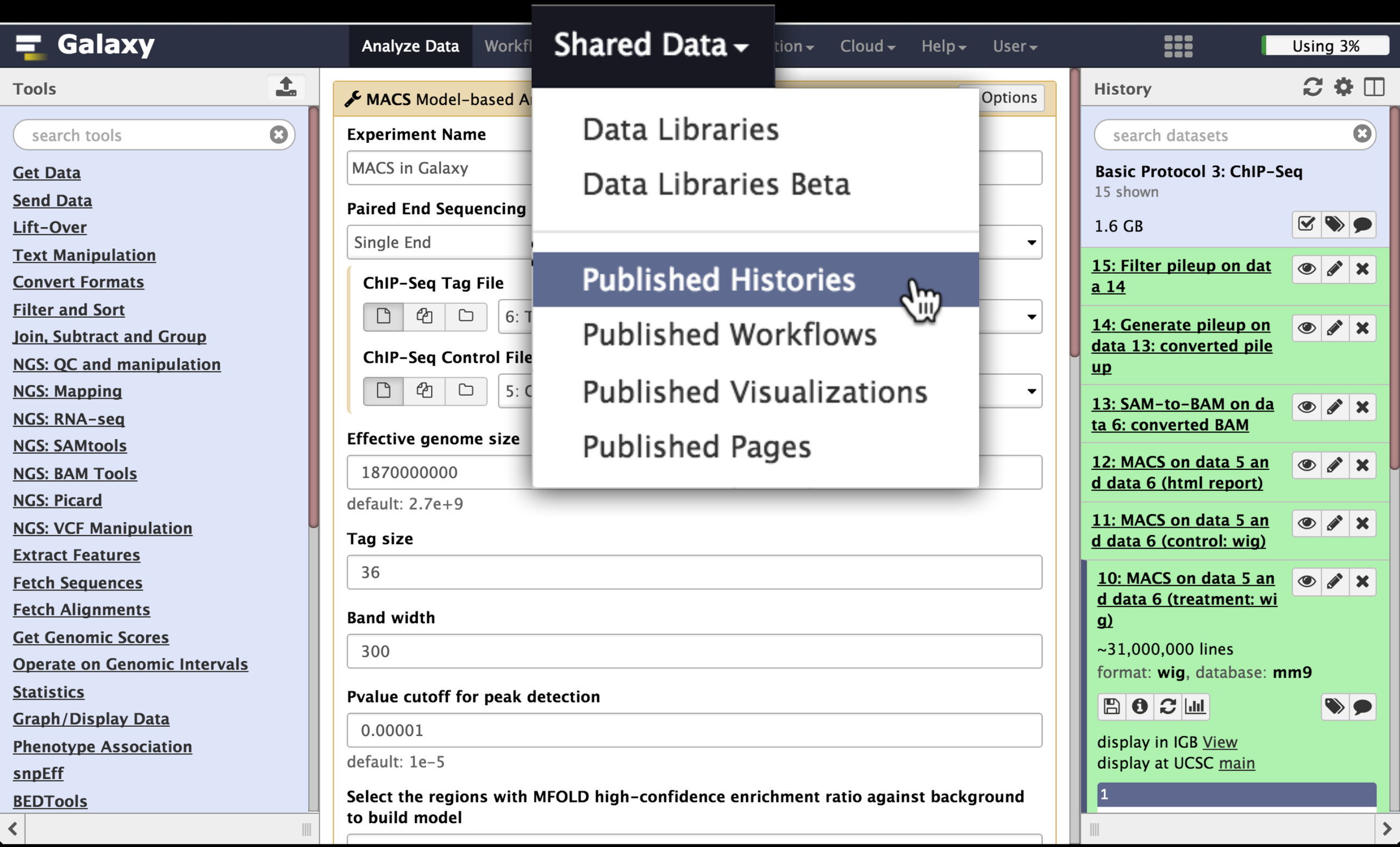
What: History / Analysis management



What: Create reusable workflows



What: Sharing and Publishing



What: Sharing and Publishing

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud | Help | User

Published Histories

search name, annotation, owner, and tags

Advanced Search

Name	Annotation	Owner	Community Rating↓	Community Tags
Infravec		dan-lawson	★★★★★	
ChIP-seq shared data		chip-seq-helin-group	★★★★★	chip illumina
Galaxy vs MEGAN	Comparison of Galaxy vs. MEGAN pipeline.	aun1	★★★★★	megan galaxy metagenomics
TRAPLINE: miRNA Targets Input	This history includes the optionally miRNA target prediction files of TRAPLINE. www.sbi.uni-rostock.de/RNAseqTRAPLINE	mwolfien	★★★★★	target prediction mirna
RNA-seq shared data		rna-seq-helin-group	★★★★★	illumina rnaseq
Galaxy Variant 101	Mother-Child mitochrondrial variation analysis. See Page https:/usegalaxy.org/u/galaxyproject/p/galaxy-101-ngs-variant	galaxyproject	★★★★★	
MOL470_Pset3 All		jbgreisman	★★★★★	
SM_1186088	Datasets correspond to our paper published in Science by Peleg et al. entitled : Altered histone acetylation is associated with age-dependent memory...	publicdata	★★★★★	
SNP Calling		jallen	★★★★★	

What: Publishing *Semantics*

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud | Help | User | Using 3%

Published Pages | mwolfien | TRAPLINE - Manual

TRAPLINE: A standardized and automated pipeline for RNA sequencing data analysis and evaluation

Markus Wolfien, Christian Rimmbach, Ulf Schmitz, Julia Jeannine Jung, Stefan Krebs, Gustav Steinhoff, Robert David, and Olaf Wolkenhauer

Department of Systems Biology and Bioinformatics, University of Rostock, 18057 Rostock, Germany

Reference und Translation Center for Cardiac Stem Cell Therapy (RTC), University of Rostock, 18057 Rostock, Germany

Gene Center Munich, LMU Munich, 81377 Munich, Germany

Stellenbosch Institute of Advanced Study (STIAS), Wallenberg Research Centre at Stellenbosch University, 7602 Stellenbosch, South Africa

Correspondence to: markus.wolfien@uni-rostock.de

We critically compare and evaluate state-of-the-art bioinformatics approaches and present a workflow that integrates the best performing data analysis and data evaluation methods in a **T**ransparent, **R**eproducible and **A**utomated **P**ipe**LINE** (**TRAPLINE**) for RNA sequencing data analysis. A comparative transcriptomics analysis with TRAPLINE results in a set of differentially expressed genes, their corresponding protein-protein interactions, a analysis of differential splicing and promoter testing and an integrated miRNA target prediction. Ultimately, the user will receive a ready-to-use file which can be importet to Cytoscape.

TRAPLINE supports NGS research by providing a workflow that requires no bioinformatics skills and decreases the processing time of the analysis.

Our pipeline is implemented in the biomedical research platform Galaxy and is freely accessible via:

Galaxy Workflow | RNAseqTRAPLINE

RNA sequencing data analysis in a Transparent Reproducible and Automated PipeLINE - TRAPLINE.

Step by Step instructions for the usage:

- **Do** your experiments (Illumina, SOLiD, Solexa Sequencing) and obtain the **FASTQ files**

Note: the analysis is predefined for the comparison of two experimental conditions with a triplicate for each experimental setup

- **Go** to the Galaxy website https://usegalaxy.org
- If you are new to Galaxy please **create** an account

About this Page

Author

mwolfien

Related Pages

All published pages

Published pages by mwolfien

Rating

Community

(2 ratings, 5.0 average)

Yours

Tags

Community:

workflow | protein_interaction | rnaseq | analysis&evaluation | mirna_prediction

Yours:

Galaxy is available several ways ...

http://galaxyproject.org

As a free for everyone service on the web: usegalaxy.org

https://usegalaxy.org

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud | Help | User | Using 3%

Tools

search tools

- Get Data
- Send Data
- Lift-Over
- Text Manipulation
- Convert Formats
- Filter and Sort
- Join, Subtract and Group
- NGS: QC and manipulation
- NGS: Mapping
- NGS: RNA-seq
- NGS: SAMtools
- NGS: BAM Tools
- NGS: Picard
- NGS: VCF Manipulation
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Phenotype Association
- snpEff
- BEDTools
- Genome Diversity
- EMBOSS

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources.

Galaxy 101

Start small

The very first tutorial you need

Tweets Follow

NIH BD2K @NIH_BD2K 1h
Submit #BD2K #Hackathon Proposals to the BD2K Centers Coord. Center! Due OCT15 Read more at ow.ly/SlUkm pic.twitter.com/2bUDJh1tJZ
Retweeted by Galaxy Project
Show Photo

Dawei Lin @iGenomics 23h
@mike_schatz My former group at UC Davis has been update an AMI with Galaxy bioinformatics.ucdavis.edu/software/
Retweeted by Galaxy Project
Expand

Tweet to @galaxyproject

PENNSTATE 1855 | JOHNS HOPKINS UNIVERSITY | TACC | iPlant Collaborative

A free for everyone web service:

http://usegalaxy.org

A free (for everyone) web server integrating a wealth of tools, compute resources, petabytes of reference data and permanent storage

TACC

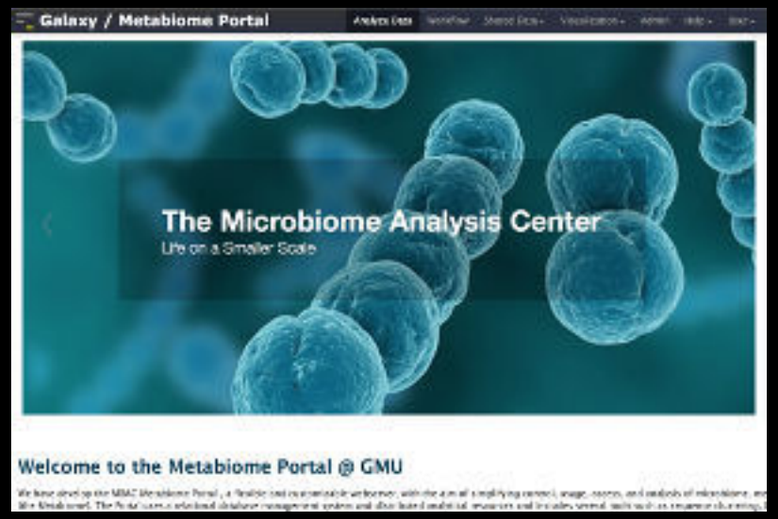
iPlant Collaborative™

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

Explore the Galaxy with

RNA-Rocket

PATHOGENPORTAL
THE BIOINFORMATICS RESOURCE CENTERS PORTAL



香港中文大學 - 華大基因跨組學創新研究院
CUHK-BGI Innovation Institute of Trans-Omics

(GIGA)^n Galaxy
by CBIIT

Integrated publishing of workflows from (GIGA)^n SCIENCE

Cistrome

A Galaxy Server dedicated to ChIP-* analysis

070+

Public Galaxy Servers
and *still* counting

The Genomic HyperBrowser

Powered by Galaxy

SCDE
STEM CELL DISCOVERY ENGINE

Experiments Connected

South Green
bioinformatics platform

Genomic analysis tools for southern and Mediterranean plants

bit.ly/gxyServers

Galaxy is available as Open Source Software

Galaxy is installed in locations around the world.

http://getgalaxy.org

Galaxy is available on the Cloud

http://aws.amazon.com/education

http://globus.org/

http://wiki.galaxyproject.org/Cloud

Galaxy on the Cloud: 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

CLOUDMAN

Why Galaxy
inside Core Facilities?

Get the advantages of Galaxy in your core

Job tracking
Analysis histories
Reusability
Reproducibility
Data Management

Galaxy API

Project Versions
latest

RTD Search
Go
Full-text doc search.

Table Of Contents
Galaxy API Documentation
Background
Quickstart
API Controllers
datasets Module
folder_contents Modul
folders Module
forms Module
genomes Module
group_roles Module
group_users Module
groups Module
histories Module
history_contents Modu
item_tags Module
libraries Module
library_contents Modu
permissions Module
quotas Module
request_types Module
requests Module
roles Module

Galaxy API Documentation

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, such as looping and branching.

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:

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

https://galaxy-dist.readthedocs.org/

Why Galaxy
for Core Facility Clients?

Empower your clients to actually use the data you generate for them **without**

learning a programming language, command line / shell interfaces, Linux package management, ...

or **extensive hand-holding** from core facility staff

Empower your clients *with Galaxy*:

Low hanging fruit

Point them at a Galaxy server for their research domain when you give them their data.

bit.ly/gxyServers

Empower your clients with Galaxy:

Moderate

Deliver data inside a Galaxy instance
with appropriate tools and reference datasets,
inside

a virtual machine image,
a Docker container,
or an Amazon Machine Image (AMI)

Empower your clients with Galaxy:

High

Deliver data inside
a core hosted, or institution hosted
Galaxy instance.

Open discussion:

What is the role of cores in supporting client data analysis?

Should this be part of your value proposition?

2016 Galaxy Community Conference (GCC2016)

June 25-29, 2016
Bloomington, Indiana

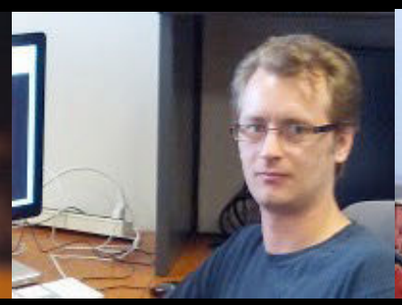
galaxyproject.org/GCC2016

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

Acknowledgements

Matt Settles
Ann Norton
Bridget the AV Guru

WACD

ABRF

NIH
Johns Hopkins University
Penn State University
Huck Institute

Thanks