

Galaxy Overview The Basics

Matt Gitzendanner
maqitz@ufl.edu
6/11/14



UF Research Computing
Information Technology
Home of High-Performance Computing and **HiPerGator**



UF Information Technology www.it.ufl.edu

UF Research Computing

UF Research Computing
Information Technology
Home of High-Performance Computing and **HiPerGator**

- ▶ Mission
 - Improve opportunities for research and scholarship
 - Improve competitiveness in securing external funding
 - Provide high-performance computing resources **and support** to UF researchers

UF Information Technology www.it.ufl.edu





UF Information Technology www.it.ufl.edu

HiPerGator

The University of Florida Supercomputer for Research


- 16,384 cores—total of about 21,000 cores today
- Infiniband interconnect
- >3PB fast, high-availability, storage
- **GPGPUs**
- Large memory (**512GB to 1TB of RAM**) nodes



UF Information Technology www.it.ufl.edu

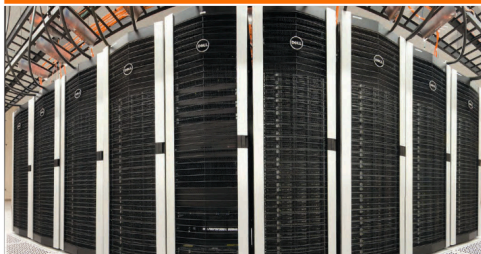
Approaches





UF Information Technology www.it.ufl.edu

UNIVERSITY OF FLORIDA | High-Performance Computing



HiPerGator


The University of Florida Supercomputer for Research

UF Information Technology www.it.ufl.edu

Cluster basics

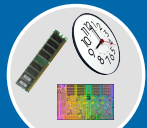
User interaction

Galaxy




Login node
(Head node)

Scheduler



Tell the scheduler
what you want to do


Compute resources



Your job runs on the cluster

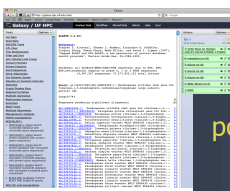
UF Information Technology www.it.ufl.edu


What is Galaxy?



Galaxy Provides Life Support for NGS Exploration

Bonus Content: Open Source





powered by
Galaxy

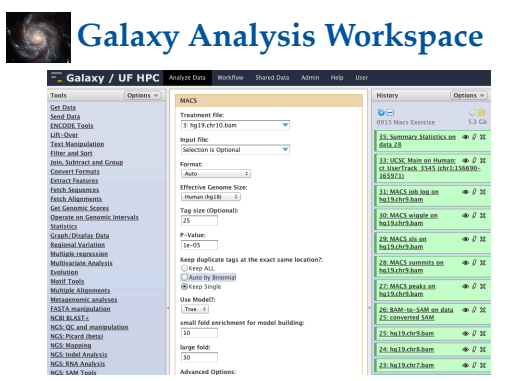
UF Information Technology www.it.ufl.edu

Galaxy: Data intensive biology for everyone

- Accessible, reproducible, transparent computational biology
- galaxy.hpc.ufl.edu
 - Local instance of Galaxy
 - Faster access to storage, easier upload
 - Local compute resources
 - Local control

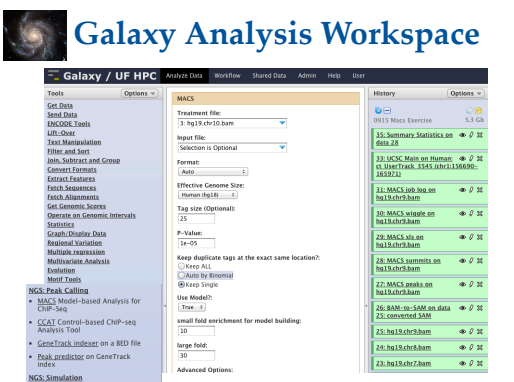
UF Information Technology www.it.ufl.edu

Galaxy Analysis Workspace



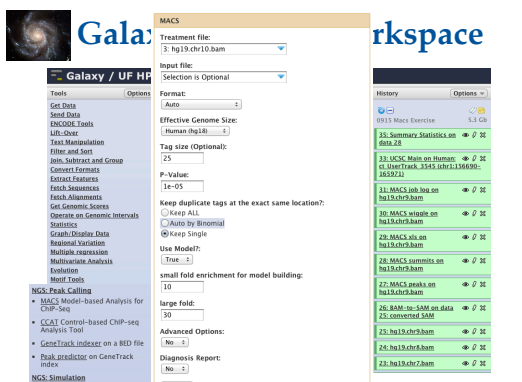
UF Information Technology www.it.ufl.edu

Galaxy Analysis Workspace



UF Information Technology www.it.ufl.edu

Galaxy Analysis Workspace



UF Information Technology www.it.ufl.edu

Galaxy / UF Information Technology

MACS

Treatment file: hg19.chr9.bam

Input file: Selection is Optional

Format: Auto

Effective Genome Size: Human (hg19)

Tag size (Optional): 25

P-Value: 1e-05

Keep duplicate tags at the exact same location: Keep ALL Write by Binomial

Use Model: True False

small fold enrichment for model building: 10

large fold: 30

Advanced Options: No

Diagnosis Report: No

Execute

History

- 0915 Macs Exercise 5.3 GB
- 35: Summary Statistics on hg19 chr9
- 33: UCSC Main on Human
- 31: UserTrack 1x5 chr11:16690-165971
- 31: MACS job log on hg19.chr9.bam
- 30: MACS wiggle on hg19.chr9.bam
- 29: MACS sig on hg19.chr9.bam
- 28: MACS summits on hg19.chr9.bam
- 27: MACS peaks on hg19.chr9.bam
- 26: BAM to SAM on data
- 25: converted SAM
- 23: hg19.chr9.bam

www.it.ufl.edu

Galaxy

chr	start	end	MACS_peak	score
chr1	179677	179078	MACS_peak_1	14.00
chr9	503365	503366	MACS_peak_2	17.00
chr9	164211	164212	MACS_peak_3	20.00
chr9	2241905	2241906	MACS_peak_4	15.00
chr9	2162896	2162897	MACS_peak_5	14.00
chr9	3467733	3467734	MACS_peak_6	14.00
chr9	3162975	3162976	MACS_peak_7	15.00
chr9	3899982	3899983	MACS_peak_8	17.00
chr9	3907658	3907659	MACS_peak_9	15.00
chr9	4315804	4315805	MACS_peak_10	17.00
chr9	4887865	4887866	MACS_peak_11	11.00
chr9	5186618	5186619	MACS_peak_12	13.00
chr9	5478013	5478014	MACS_peak_13	14.00
chr9	5515840	5515841	MACS_peak_14	13.00
chr9	5564231	5564232	MACS_peak_15	11.00
chr9	5695455	5695456	MACS_peak_16	9.00
chr9	5822436	5822439	MACS_peak_17	12.00
chr9	6038019	6038020	MACS_peak_18	16.00
chr9	6481221	6481222	MACS_peak_19	29.00
chr9	6757871	6757872	MACS_peak_20	12.00
chr9	7028374	7028375	MACS_peak_21	11.00
chr9	9428809	9428810	MACS_peak_22	8.00
chr9	9442235	9442236	MACS_peak_23	5.00
chr9	9487422	9487423	MACS_peak_24	3.00
chr9	9524985	9524986	MACS_peak_25	5.00
chr9	9677411	9677412	MACS_peak_26	7.00
chr9	1277444	1277447	MACS_peak_27	14.00
chr9	1302479	1302479	MACS_peak_28	12.00
chr9	1420262	1420263	MACS_peak_29	12.00
chr9	1503446	1503447	MACS_peak_30	7.00
chr9	1670476	1670477	MACS_peak_31	10.00
chr9	1670476	1670477	MACS_peak_32	10.00
chr9	1694119	1694120	MACS_peak_33	12.00
chr9	1705070	1705071	MACS_peak_34	11.00
chr9	1766373	1766374	MACS_peak_35	10.00
chr9	1814892	1814893	MACS_peak_36	9.00
chr9	1890955	1890955	MACS_peak_37	11.00
chr9	21085743	21085742	MACS_peak_38	47.00
chr9	2219180	2219180	MACS_peak_39	16.00
chr9	2201638	2201639	MACS_peak_40	7.00
chr9	2201638	2201639	MACS_peak_41	7.00

www.it.ufl.edu

Galaxy

Metadata

History: LANA CHIP peaks on hg19 5.3 GB

Tags: LANA, chip, hg19, peaks, chr9

Annotation / Notes: Peak calling on LANA CHIP-seq data using Human chromosome 9 from hg19 build

27: MACS peaks on hg19.chr9.bam

236 regions

format: bed, database: ?

Tags: LANA, chip, hg19, chr9, MACS

view in GeneTrack

1. chrom	2. Start	3. End	4. Name
chr9	176690	179457	MACS_pea
chr9	502264	506252	MACS_pea
chr9	763181	765291	MACS_pea
chr9	2241428	2243431	MACS_pea
chr9	3161298	3162300	MACS_pea
chr9	3467312	3468066	MACS_pea

www.it.ufl.edu

Getting Data into Galaxy

- Upload a file from your computer
 - Direct upload (<2GB)
 - For large files: scp or copy files to HPC
 - Load from within Galaxy
 - http://wiki.hpc.ufl.edu/index.php/Galaxy_Data_Import
- External data
 - UCSC table browser
 - Biomart
 - interMine / modMine
 - EuPathDB
 - EncodeDB
 - EpiGRAPH
 - FlyMine
 - GraveneMart...

www.it.ufl.edu

Galaxy

Data Libraries

Data Library "GMS 6001 MACS Exercise"

MACS test data

Name	Message	Uploaded By	Date	File Size
2010-12-14_7_macs_job_log.bam		om@hpc.ufl.edu	2011-09-13	1.8 GB
2010-12-14_7_macs_job_log.sorted.bam		om@hpc.ufl.edu	2011-09-13	1.4 GB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	80.8 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	82.5 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	74.9 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	50.9 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	36.3 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	48.1 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	55.9 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	64.3 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	33.5 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	39.6 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	148.3 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	85.7 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	173 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	16.9 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	122.1 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	448.0 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	114.0 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	85.7 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	102.7 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	63.7 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	89.9 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	85.9 MB
h31.chr11.bam		om@hpc.ufl.edu	2011-09-14	64.8 MB

www.it.ufl.edu

Galaxy

Data Access Control

Roles associated with new group: HPC test ChIP-seq analyses

Groups:

Name	Users	Role
HPC	0	2
Taylor HPC Lab	2	1

Roles:

Name	Description	Type	Groups
HPC	Role for group HPC	system	1
HPC-test ChIP-seq analyses	Test analyses of ChIP-seq data	admin	1

Users:

Email	User Name	Groups	Roles	External	Last Login
adrian@ufl.edu	adrian	0	1	yes	Sep 15, 2011
bosnick@ufl.edu	bosnick	0	1	yes	Sep 15, 2011
cpaves@ufl.edu	cpaves1	0	1	yes	Sep 15, 2011
ceffrey@ufl.edu	ceffrey	0	1	yes	Sep 15, 2011
collis@ufl.edu	collis3	0	1	yes	Sep 15, 2011

www.it.ufl.edu

Galaxy Tool Suites

- ▶ Text Manipulation
- ▶ Format Converters
- ▶ Filtering and Sorting
- ▶ Join, Subtract, Group
- ▶ Sequence Tools
- ▶ Multi-species Alignment Tools
- ▶ Genomic Interval Operation
- ▶ Summary Statistics, graphing

- ▶ Regional Variation
- ▶ EMBOSS
- ▶ Evolution
- ▶ RNA-Seq
- ▶ ChIP-Seq
- ▶ GATK
- ▶ Phylogenetics

Information Technology
 www.it.ufl.edu

A galaxy of tools

ILLUMINA DATA

FASTQ Converter: convert between various FASTQ quality formats

FASTQ subset: on joined paired end reads

FASTQ subset: 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

ABI-SOLID DATA

Convert SOLID output to fastq

Compute quality statistics for SOLID data

Draw quality score boxplots for SOLID data

EMBOSS

Metagenomic analyses

Human Genome Statistics

EMBOSS

NCBI TOOLBOX BETA

NCBI TOOLBOX BETA

NGS QC and manipulation

NGS Mapping

- 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 plots: from BAM dataset
- Filter: align on coverage and SQ
- Phrap-to-Interval: condenses phrap format into ranges of bases
- Repeat: provides simple stats on SAM files

GENETICS

GENETICS

SNP/WGA: Data Filters

SNP/WGA: QC: LID: Peaks

SNP/WGA: Statistical Models

NGS SAM Tools

NGS: Index Analysis

- Filter: Index for SAM
- Extract: Index from SAM

NGS: Peak Calling

- MACS: Model-based Analysis of ChIP-Seq
- GeneTrack: Index on a BED file
- Peak analysis: on GeneTrack index

RNA-Seq

- Tophat: Find splice junctions using RNA-seq data
- Cuffdiff: transcript assembly and FPKM (RPKM) estimates for RNA-seq data
- Cuffmerge: compare assembled transcripts to a reference annotation and track Cuffdiff annotations across multiple experiments
- Cuffdiff2: find significant changes in transcript expression, splicing, and promoter use

FILTERING

- Filter: Combine Transcripts using tracking file

Information Technology
 www.it.ufl.edu

Galaxy Workflows

Unknown

This tool cannot be used in workflow

BAM-to-SAM

Include "BAM-to-SAM" in workflow

Convert Genomic Intervals To Strict BEDs

Include "Convert Genomic Intervals To Strict BEDs" in workflow

MACS

Include "MACS" in workflow

Convert BED to GeneTrack Index

Include "Convert BED to GeneTrack Index" in workflow

25: hg19.chr9.bam	▶	Deduct Workflow Dataset Security	
26: BAM-to-SAM on data 25: converted SAM	▶	Show Deleted Datasets	
27: MACS peaks on hg19.chr9.bam	▶	Show Structure	
28: MACS summits on hg19.chr9.bam	▶	Export to File	
29: MACS xls on hg19.chr9.bam	▶	Delete	
30: MACS wiggle on hg19.chr9.bam	▶	Other Actions	
31: MACS job log on hg19.chr9.bam	▶	Import from File	

Information Technology
 www.it.ufl.edu

Galaxy Workflows

Workflow Canvas: Workflow constructed from history 'LANA ChIP peaks on hg19'

Details

Tool: MACS

Treatment File: Data input 'file' interval or sam or bam or fland or elandmulti or bed

Input File: Data input 'file' interval or sam or bam or fland or elandmulti or bed

Format: Auto

Effective Genome Size: Human (hg19)

Tag size (Optional): 25

Edit Workflow Attributes

Name: Workflow constructed from history 'LANA ChIP peaks on hg19'

Tags: LANA x | ChIP-Seq x | hg19 x

Apply tags to make it easy to search for and find items with the same tag

Annotation / Notes: This is a partial peak calling with MACS using hg19 and chr9 data

Information Technology
 www.it.ufl.edu

Galaxy Workflows

Information Technology
 www.it.ufl.edu

Sharing and publishing

Share or Publish History 'LANA ChIP peaks on hg19'

Making History Accessible via Link and Publishing It

This history is currently restricted so that only you and the users listed below can access it. You can:

- Make History Accessible via Link**
Generates a web link that you can share with other people so that they can view and import the history.
- Make History Accessible and Publish**
Makes the history accessible via link (see above) and publishes the history to Galaxy's **Published Histories** section, where it is publicly listed and searchable.

Sharing History with Specific Users

You have not shared this history with any users.

Share with a user

[Back to Histories List](#)

Information Technology
 www.it.ufl.edu

Sharing and publishing

Share or Publish History 'LANA ChIP peaks on hg19'

Making History Accessible via Link and Publishing It
This history is currently accessible via link and published.
Anyone can view and import this history by visiting the following URL:
<http://galaxy.hpc.ufl.edu/ui/meskalenko/h/ana-chip-peaks-on-hg19>
This history is publicly listed and searchable in Galaxy's **Published Histories** section.
You can:

- Unpublish History**
Removes this 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 this 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
The following users will see this history in their history list and will be able to view, import, and run it.

Email
magatz@ufl.edu

Share with another user

UF Information Technology www.it.ufl.edu

Summary

- Analyze data without the CLI
- Visualize the results
- Publish histories, workflows, and annotated pages
- Add new tools, get support @ HPC
- Focus on your science, not minutiae
- UF Galaxy** – coming to a browser near you!

UF Information Technology www.it.ufl.edu

Demo

UF Information Technology www.it.ufl.edu

Galaxy demo

<http://galaxy.hpc.ufl.edu>

UF Information Technology www.it.ufl.edu

UF Research Computing

- Help and Support
 - Help Request Tickets
 - <https://support.hpc.ufl.edu>
 - For any kind of question or help requests
 - <http://wiki.hpc.ufl.edu>
 - Documents on hardware and software resources
 - Various user guides
 - Many sample submission scripts
 - <http://hpc.ufl.edu>
 - Frequently Asked Questions
 - Account set up and maintenance

UF Information Technology www.it.ufl.edu