

# Introduction to Galaxy

Matt Gitzendanner magitz@ufl.edu  
Oleksandr Moskalenko om@hpc.ufl.edu

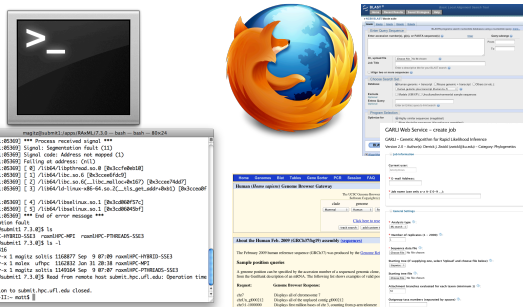
UF Information Technology [www.it.ufl.edu](http://www.it.ufl.edu)

## Today's research computing



UF Information Technology [www.it.ufl.edu](http://www.it.ufl.edu)



## Approaches



UF Information Technology [www.it.ufl.edu](http://www.it.ufl.edu)

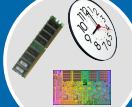
## Cluster basics

User interaction


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](http://www.it.ufl.edu)

## What is Galaxy?



UF Information Technology [www.it.ufl.edu](http://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](http://www.it.ufl.edu)

## Galaxy Analysis Workspace

## Galaxy Analysis Workspace

## Galaxy Analysis Workspace

## Galaxy Analysis Workspace

## Galaxy Analysis Workspace

## Galaxy Analysis Workspace

### Metadata

## 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
  - GrameneMart...

UF Information Technology | www.it.ufl.edu

## Data Libraries

Data Library "GMS 6001 MACS Exercise"

Name	Message	Uploaded By	Date	File Size
2010-12-14 7_16133_ahp_sorted.bam		omihpc@ufl.edu	2011-09-15	1.8 GB
2010-12-14 7_16133_ahp_sorted.bam		omihpc@ufl.edu	2011-09-15	1.4 GB
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	80.8 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	82.5 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	74.9 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	50.9 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	36.1 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	48.1 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	55.9 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	64.5 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	33.5 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	28.6 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	145.5 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	38.5 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	17.5 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	16.9 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	126.3 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	448.0 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	118.0 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	85.7 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	102.7 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	67.8 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	89.9 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	65.9 Mb
hg19.chr11.bam		omihpc@ufl.edu	2011-09-14	64.8 Mb

UF Information Technology | www.it.ufl.edu

## Data Access Control

Roles associated with new group  
HPC test CHIP-seq analyses

Name	Users	Role
HPC	0	2
Taylor HPC Lab	2	1

Users associated with new group

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
adelson@ufl.edu	adelson	0	1	yes	Sep 15, 2011
bozwick@ufl.edu	bozwick	0	1	yes	Sep 15, 2011
cupress@ufl.edu	cupress	0	1	yes	Sep 15, 2011
cliffrey@ufl.edu	cliffrey	0	1	yes	Sep 15, 2011
coltr@ufl.edu	coltr	0	1	yes	Sep 15, 2011

UF Information Technology | 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

UF Information Technology | www.it.ufl.edu

## A galaxy of tools

<b>CL, QC and manipulation</b> ILLUMINA DATA FASTQ Converter FASTQ quality filters FASTQ splitter FASTQ joiner FASTQ Summary Statistics by column RICHIE-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 heatmap for SOLID data GENOMIC FASTQ MANIPULATION Filter FASTQ reads by quality score and length FASTQ Trimmer by column FASTQ Quality Trimmer by sliding window	<b>Metagenomic analyses</b> Human Genome Variation EMBOSS NGS TOOLBOX BETA <b>NGS, QC and manipulation</b> NGS Mapping Map with Bowtie for Illumina Map with BWA for Illumina RICHIE-454 Lazy map short reads against reference sequence Mapblast compare short reads against tips, nt, and wgs databases ABI-SOLID Parse Blast XML output Convert SOLID output to Fastq Map with Bowtie for SOLID <b>NGS, SAM Tools</b> NGS: Index Analysis NGS: Peak Calling NGS: RNA Analysis BEDTOOLS SNP/WGA: Data Filters SNP/WGA: QC: LD: Plots SNP/WGA: Statistical Models	<b>NGS TOOLBOX BETA</b> <b>NGS, QC and manipulation</b> 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 atlas from BAM dataset Filter atlas on coverage and QV Filter-to-interval condenses group format into ranges of bases Mapping provides simple stats on BAM files <b>NGS: Index Analysis</b> NGS: Peak Calling NGS: RNA Analysis BEDTOOLS SNP/WGA: Data Filters SNP/WGA: QC: LD: Plots SNP/WGA: Statistical Models	<b>NGS, SAM Tools</b> NGS: Index Analysis Filter indices for SAM Extract indices from SAM Index Analysis NGS: Peak Calling MGS Model-based Analysis of ChIP-Seq GeneTrack: Index on a BED file Peak predictor on GeneTrack index <b>NGS: RNA Analysis</b> RNA-SEQ TopHat Find splice junctions using RNA-seq data Cuffdiff transcript assembly and FPKM/RPM estimates for RNA-Seq data Cuffdiff compare assembled transcripts to a reference annotation and track Cuffdiff transcripts across multiple experiments Cuffdiff find significant changes in transcript expression Filter Filter Combined Transcripts using tracking file
---	--	---	--

UF Information Technology | www.it.ufl.edu

## Galaxy Workflows

Unknown This tool cannot be used in workflows BAM-to-SAM Include "BAM-to-SAM" in workflow	25: hg19.chr9.bam Treat as input dataset 26: BAM-to-SAM on data 25: converted SAM 27: MACS peaks on hg19.chr9.bam 27: MACS peaks on hg19.chr9.bam 28: MACS sums on hg19.chr9.bam 29: MACS xls on hg19.chr9.bam 30: MACS wiggle on hg19.chr9.bam 31: MACS job log on hg19.chr9.bam Convert BED to GeneTrack: Index Include "Convert BED to GeneTrack Index" in workflow 27: MACS peaks on hg19.chr9.bam	Extract Workflow Dataset Security Show Deleted Datasets Show Hidden Datasets Show Structure Export to File Delete Other Actions Import from File 27: MACS peaks on hg19.chr9.bam 26: BAM-to-SAM on data 25: converted SAM 25: hg19.chr9.bam 24: hg19.chr9.bam 23: hg19.chr9.bam 22: hg19.chr9.bam 21: hg19.chr9.bam 20: hg19.chr9.bam
--	---	---

UF 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 'tfile' (interval or sam or bam or fastq or readmulti or bed)

Input file: Data input 'tfile' (interval or sam or bam or fastq or readmulti or bed)

Format:

Effective Genome Size: (Human hg19)

Tag size (Optional):

**Summary Statistics**

Summary statistics on:

**Details**

**Edit Workflow Attributes**

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

Tags:

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

UF Information Technology | www.it.ufl.edu

## Galaxy Workflows

UF 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](#)

UF 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/~moskalenko/lana-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**

**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

**Galaxy / UF HPC** | Analyze Data | Workflow | Shared Data | Help | User

**Tools**

- Get Data
- Send Data
- ENCODE Tools
- Life-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Model Tools
- Multiple Alignments
- Metagenomic analyses
- FASTA manipulation
- NCR-BLAST
- NCR-DK and manipulation

**History**

LANA ChIP peaks on hg19

0 bytes

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

**UF**  
UNIVERSITY OF FLORIDA

**UF HPC Galaxy News:**

- 2001-08-09: Prototype Galaxy Instance

An instance of Galaxy Platform for Biological Research Computing was brought online at the University of Florida High-Performance Computing Center for testing and demonstration purposes. This instance is not available for public use, yet. However, you can email HPC or the biological applications support directly to request to be notified of its general availability.

The Galaxy project is supported in part by NSF, NH&I, and the Huck Institutes of the Life Sciences.

UF Information Technology | www.it.ufl.edu

## Galaxy demo

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

### UF HPC Center Login

Username:

Password:

[Request an account](#)  
[Reset my password](#)

## UF Research Computing

- ▶ Help and Support
  - Help Request Tickets
    - <https://support.hpc.ufl.edu>
    - For any kind of question or help requests
    - Searchable database of solutions
  - We are here to help!
    - [support@hpc.ufl.edu](mailto:support@hpc.ufl.edu)



## Training Schedule

- ✓ Aug 28: Intro to UFHPC, getting started
- ✓ Sept 10: Modules, RHEL6 Transition, User Q&A
- ✓ Sept 17: The Linux/Unix Shell - An Introduction
- ✓ Sept 24: Running Jobs, Submission Scripts, Modules
- ✓ Oct 1: Galaxy Overview, The Basics
- ▶ Oct 8: NGS Data Techniques: General Methods and Tools
- ▶ Oct 15: NGS Data Techniques: Reference Based Mapping and de Novo Assembly
- ▶ Oct 22: Phylogenetic Analyses
- ◆ Oct 29: Research Computing Day: Moving Big Data
- ▶ Nov 5: Multiprocessing at the HPC Center
- ▶ Nov 12: Using Git and CMake to Organize and Drive Data Analysis Pipelines
- ▶ Nov 19: Introduction to GPU Nodes
- ▶ Nov 29: NGS Data Techniques: RNA-Seq
- ▶ Dec 3: NGS Data Techniques: Alternative Splicing Analysis

## UF Research Computing

- ▶ Help and Support (Continued)
  - <http://wiki.hpc.ufl.edu>
    - Documents on hardware and software resources
    - Various user guides
    - Many sample submission scripts
  - <http://hpc.ufl.edu/support>
    - Frequently Asked Questions
    - Account set up and maintenance

