



Galaxy For Users

Oleksandr Moskalenko, Ph.D.
University of Florida Research Computing

Super Computing 2012, Salt Lake City, Nov 12-15, 2012

What is Galaxy?



Nov 12-15, 2012

SC12

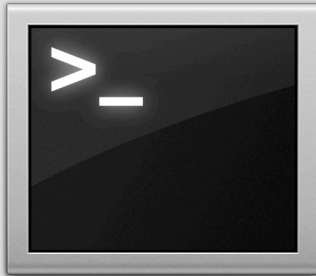
Computational biology **platform**

- Open
- Web-based
- Accessible
- Reproducible
- Transparent
- Flexible
- Secure
- Collaborative



Batch Environment

Head node



Login to
head
node

Scheduler



Interactive
session or
batch
submission

Computing
resources



Your job
runs on the
cluster

Galaxy Analysis Workspace

The screenshot displays the Galaxy Analysis Workspace interface. At the top, the navigation bar includes "Galaxy / UF HPC" and menu items: "Analyze Data", "Workflow", "Shared Data", "Admin", "Help", and "User".

The main workspace is divided into three panels:

- Tools Panel (Left):** A list of tool categories including "Get Data", "Send Data", "ENCODE Tools", "Lift-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", "Motif Tools", "Multiple Alignments", "Metagenomic analyses", "FASTA manipulation", "NCBI BLAST+", "NGS: QC and manipulation", "NGS: Picard (beta)", "NGS: Mapping", "NGS: Indel Analysis", "NGS: RNA Analysis", and "NGS: SAM Tools".
- MACS Tool Configuration (Center):**
 - Treatment file:** 3: hg19.chr10.bam
 - Input file:** Selection is Optional
 - Format:** Auto
 - Effective Genome Size:** Human (hg18)
 - Tag size (Optional):** 25
 - P-Value:** 1e-05
 - Keep duplicate tags at the exact same location?:** Keep Single
 - Use Model?:** True
 - small fold enrichment for model building:** 10
 - large fold:** 30
 - Advanced Options:** (collapsed)
- History Panel (Right):** A list of previous jobs with details like "0915 Macs Exercise" (5.3 Gb) and a list of jobs including "35: Summary Statistics on data 28", "33: UCSC Main on Human: ct UserTrack 3545 (chr1:156690-165971)", "31: MACS job log on hg19.chr9.bam", "30: MACS wiggle on hg19.chr9.bam", "29: MACS xls 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", "25: hg19.chr9.bam", "24: hg19.chr8.bam", and "23: hg19.chr7.bam".



Tool panel

The screenshot displays the Galaxy / UF HPC interface. The top navigation bar includes 'Galaxy / UF HPC', 'Analyze Data', 'Workflow', 'Shared Data', 'Admin', 'Help', and 'User'. The left sidebar lists various tool categories: 'Tools', 'Get Data', 'Send Data', 'ENCODE Tools', 'Lift-Over', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Convert Formats', and 'Extract Features'. The main panel shows the configuration for the 'MACS' tool. The 'Treatment file' is set to '3: hg19.chr10.bam', the 'Input file' is 'Selection is Optional', and the 'Format' is 'Auto'. Below these are fields for 'Window Size' (set to 8) and 'Optional' parameters. A checkbox 'Use tags at the exact same location?' is visible. The 'Advanced Options' section is partially visible at the bottom. On the right, the 'History' panel shows a list of recent jobs, including '0915 Macs Exercise' (5.3 Gb) and a series of jobs numbered 23 to 35, such as '35: Summary Statistics on data 28', '33: UCSC Main on Human: ct UserTrack 3545 (chr1:156690-165971)', '31: MACS job log on hg19.chr9.bam', '30: MACS wiggle on hg19.chr9.bam', '29: MACS xls 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', '25: hg19.chr9.bam', '24: hg19.chr8.bam', and '23: hg19.chr7.bam'. Each history entry includes an eye icon, a refresh icon, and a delete icon.

NGS: Peak Calling

- MACS Model-based Analysis for ChIP-Seq
- CCAT Control-based ChIP-seq Analysis Tool
- GeneTrack indexer on a BED file
- Peak predictor on GeneTrack index

NGS: Simulation

SNP/WGA: Data; Filters

NGS: SAM Tools

Advanced Options:



Tool Interface

Galaxy / UF H

Tools Option

- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features

NGS: Peak Calling

- MACS Model-based Analysis of ChIP-Seq
- CCAT Control-based ChIP Analysis Tool
- GeneTrack indexer on a BAM
- Peak predictor on GeneTrack index

NGS: Simulation

SNP/WGA: Data; Filters

NGS: SAM Tools

MACS

Treatment file:
3: hg19.chr10.bam

Input file:
Selection is Optional

Format:
Auto

Effective Genome Size:
Human (hg18)

Tag size (Optional):
25

P-Value:
1e-05

Keep duplicate tags at the exact same location?:
 Keep ALL
 Auto by Binomial
 Keep Single

Use Model?:
True

small fold enrichment for model building:
10

large fold:
30

Advanced Options:
No

Diagnosis Report:
No

Execute

User

History Options

0915 Macs Exercise 5.3 Gb

- 35: Summary Statistics on data 28
- 33: UCSC Main on Human: ct UserTrack 3545 (chr1:156690-165971)
- 31: MACS job log on hg19.chr9.bam
- 30: MACS wiggle on hg19.chr9.bam
- 29: MACS xls 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
- 25: hg19.chr9.bam
- 24: hg19.chr8.bam
- 23: hg19.chr7.bam







History Panel

The screenshot displays the Galaxy web interface. On the left is a 'Tools' sidebar with categories like 'Get Data', 'Text Manipulation', and 'Statistics'. The main area shows the 'MACS' tool configuration with fields for 'Treatment file', 'Input file', 'Format', 'Effective Genome Size', 'Tag size', 'P-Value', and 'Keep duplicate tags'. On the right, the 'History' panel lists recent jobs, including '0915 Macs Exercise' (5.3 Gb) and various analysis tasks like 'Summary Statistics on data 28', 'UCSC Main on Human: ct UserTrack 3545', and 'MACS job log on hg19.chr9.bam'.



Metadata


History Options ▾




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	502364	506252	MACS_pea
chr9	763181	765291	MACS_pea
chr9	2241428	2243431	MACS_pea
chr9	3161298	3162300	MACS_pea
chr9	3467312	3468066	MACS_pea



Data Access and Sharing

Roles associated with new group

HPC test CHIP-seq analyses

>>

Users associated with new

om@hpc.ufl.edu
magitz@ufl.edu

Role Groups

gal search

gwi Advanced Search

<input type="checkbox"/> Name ↓	Users	Roles
<input type="checkbox"/> HPC ▾	0	2

Roles

search

Advanced Search

<input type="checkbox"/> Name ↓	Description	Type	Groups
<input type="checkbox"/> HPC ▾	Role for group HPC	system	1
<input type="checkbox"/> HPC test CHIP-seq analyses ▾	Test analyses of CHIP-seq data	admin	1

Users

search

Advanced Search

<input type="checkbox"/> Email ↓	User Name	Groups	Roles	External	Last Login
<input type="checkbox"/> aedison@ufl.edu ▾	aedison	0	1	yes	Sep 15, 2011
<input type="checkbox"/> bostwick@ufl.edu ▾	bostwick	0	1	yes	Sep 15, 2011
<input type="checkbox"/> cgraves3@ufl.edu ▾	cgraves3	0	1	yes	Sep 15, 2011
<input type="checkbox"/> cjeffrey@ufl.edu ▾	cjeffrey	0	1	yes	Sep 15, 2011
<input type="checkbox"/> colltd3@ufl.edu ▾	colltd3	0	1	yes	Sep 15, 2011



Data Inflow

- UCSC table browser
- Biomart
- interMine / modMine
- EuPathDB
- EncodeDB
- EpiGRAPH
- FlyMine
- GrameneMart...
- SCP
- FTP
- Web upload



Data Outflow

- UCSC table browser
- Download
- Publish via “Pages”
- Share within Galaxy
- Import/Export between instances
- Export to other tools



Data Sharing

Share or Publish History 'LANA CHIP peaks on hg19'

Making History Accessible via Link and

This history is currently restricted so that only you

Make History Accessible via Link

Generates a web link that you can share with others

Make History Accessible and Publish

Makes the history accessible via link (see above) 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](#)

Published Pages | aun1 | Windshield Splatter

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

SERGEI KOSAKOVSKY POND^{1,2,*}, SAMIR WADHAWAN^{3,6*}, FRANCESCA CHIAROMONTE⁴, GURUPRASAD ANANDA^{1,3}, WEN-YU CHUNG^{1,3,7}, JAMES TAYLOR^{1,5}, ANTON NEKRUTENKO^{1,3} and THE GALAXY TEAM^{1,4}

Correspondence should be addressed to SK, FW, 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
Generic workflow for performing a metagenomic analysis on NGS data.

Accessing the Data

Windshield Splatter datasets analyzed in this manuscript can be accessed through this [Galaxy Library](#). From there they can be re-analyzed through Galaxy using the above workflows or downloaded.

Supplemental Analysis

About this Page

Author
aun1

Related Pages
[All published pages](#)
[Published pages by aun1](#)

Rating
Community (6 ratings, 5.0 average) ★★★★★

Tags
Community:
megan galaxy paper

Published Workflows

search name, annotation, owner, and tag

[Advanced Search](#)

Name	Annotation	Owner	Community Rating	Community Tags	Last Updated ↓
LANA CHIP peaks on hg19		moskalenko	★★★★★		2 minutes ago



Automation via workflows

The screenshot displays a workflow configuration interface. On the left, several tool panels are visible, each with a title and a checkbox for including the tool in the workflow:

- Unknown**: *This tool cannot be used in workflows*
- BAM-to-SAM**: Include "BAM-to-SAM" in workflow
- Convert Genomic Intervals To Strict BED6**: Include "Convert Genomic Intervals To Strict BED6" in workflow
- MACS**: Include "MACS" in workflow
- Convert BED to GeneTrack Index**: Include "Convert BED to GeneTrack Index" in workflow

On the right, a list of workflow steps is shown, each with a name and a right-click context menu. The steps are:

- 25: hg19.chr9.bam (with 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 summits 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
- 27: MACS peaks on hg19.chr9.bam

The context menu for step 27 is open, showing the following options:

- Extract Workflow
- Dataset Security
- Show Deleted Datasets
- Show Hidden Datasets
- Show Structure
- Export to File
- Delete
- Other Actions
- Import from File

The dataset list on the right side of the interface includes:

- 31: M hg19...
- 30: M hg19...
- 29: M hg19...
- 28: M hg19...
- 27: MACS peaks on hg19.chr9.bam
- 26: BAM-to-SAM on data 25: converted SAM
- 25: hg19.chr9.bam
- 24: hg19.chr8.bam
- 23: hg19.chr7.bam
- 22: hg19.chr6.bam
- 21: hg19.chr5.bam
- 20: hg19.chr4.bam
- 19: hg19.chr3.bam



Workflow editing

The screenshot displays a workflow editor interface. The main area is a grid-based canvas titled "Workflow Canvas | Workflow constructed from history 'LANA ChIP peaks on hg19'". It contains several workflow steps: "Input dataset" (multiple instances), "BAM-to-SAM", "Convert BED to GeneTrack Index", "MACS", and "Summary Statistics". Each step has a close button (⌘) and a refresh button (↺). The "MACS" step is selected, and its details are shown in a panel on the right.

Details Panel:

- Tool:** MACS
- Treatment file:** Data input 'tfile' (interval or sam or bam or eland or elandmulti or bed)
- Input file:** Data input 'cfile' (interval or sam or bam or eland or elandmulti or bed)
- Format:** Auto
- Effective Genome Size:** Human (hg19)
- Tag size (Optional):** 25
- P-Value:** 1e-05
- Keep duplicate tags at the exact same location?:** Keep Single
- Use Model?:** True
- small fold enrichment for model building:** 10
- large fold:** 20

Edit Workflow Attributes Panel:

- Name:** Workflow constructed from history 'LANA ChIP peaks on hg19'
- Tags:** LANA, CHIP-Seq, hg19, chr9
- Annotation / Notes:** This is a partial peak calling with MACS using hg19 and chr9 data



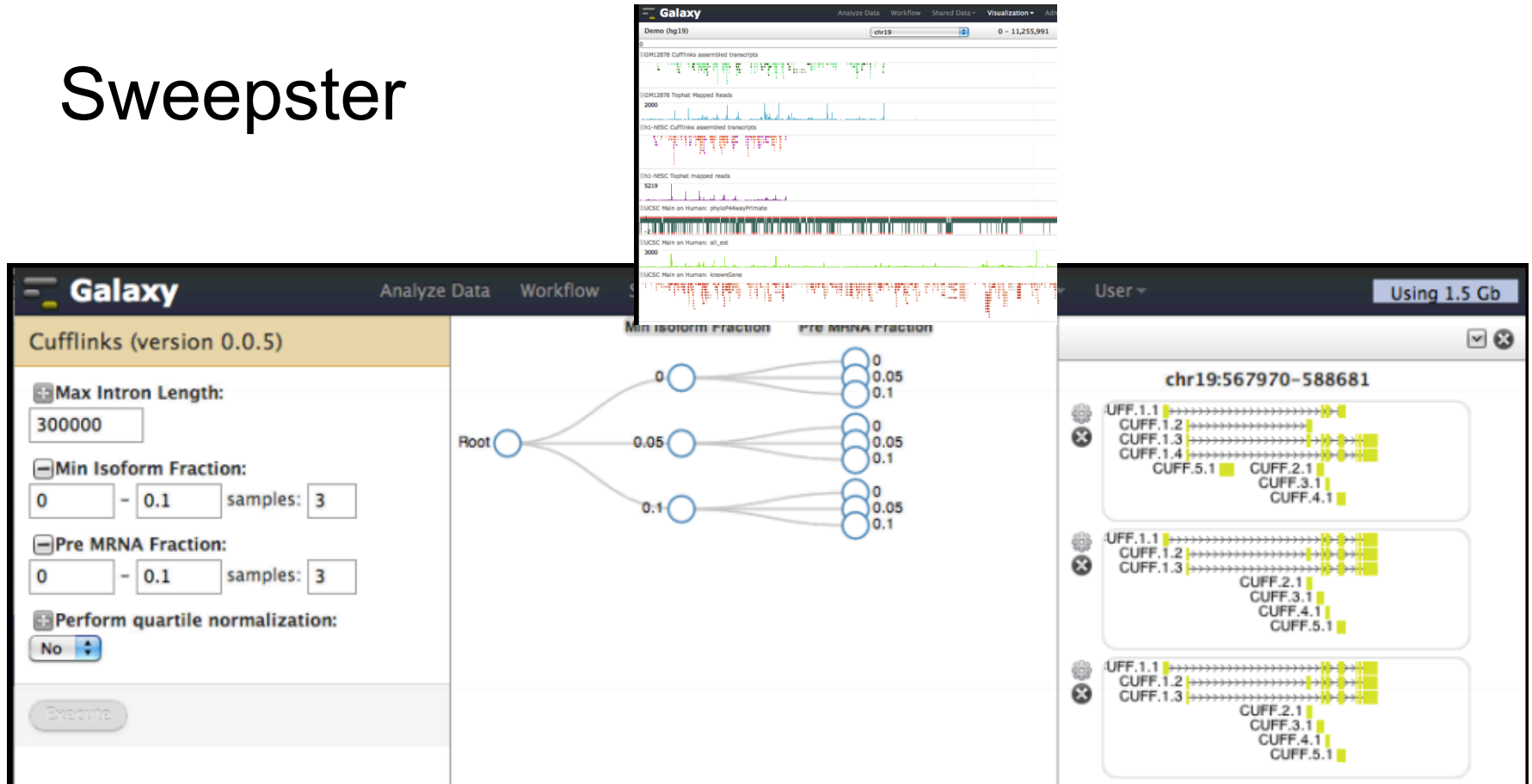
Visual Interactive Analysis

Trackster



Visual Interactive Analysis

Sweepster



Administrative side

- User, data library, job administration, management and reporting
- Integration of new tools via wrappers
- Modular and scalable (database, web server, web and job runners)
- API for command-line access
- ToolShed model for CLI tools, ENV



Batch System Integration

- All major schedulers
- Condor; XSEDE
- drmaa and pbs_python libraries
- Fixed and dynamic job runners
- Free-for-all and real user jobs
- Infinite scaling



Community

- Large and growing community
- Mailing lists
- Revision control system
- Issue tracking
- IRC
- Community Conference



Summary

- Web-based computational framework
- Future of research computing
- Built for researchers
- Flexible and highly functional



Galaxy Team



Enis Afgan
IRB



Guru Ananda
Penn State



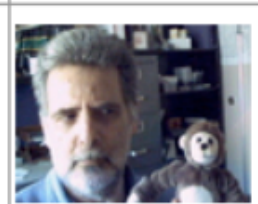
Dannon Baker
Emory



Jennifer Hillman Jackson
Penn State



Greg von Kuster
Penn State



Ross Lazarus
Harvard & BakerIDI



Dan Blankenberg
Penn State



Dave Bouvier
Penn State



Dave Clements
Emory



Rémi Marenco
Emory



Scott McManus
Emory



Anton Nekrutenko
Penn State



Nate Coraor
Penn State



Carl Eberhard
Emory



Jeremy Goecks
Emory



James Taylor
Emory

