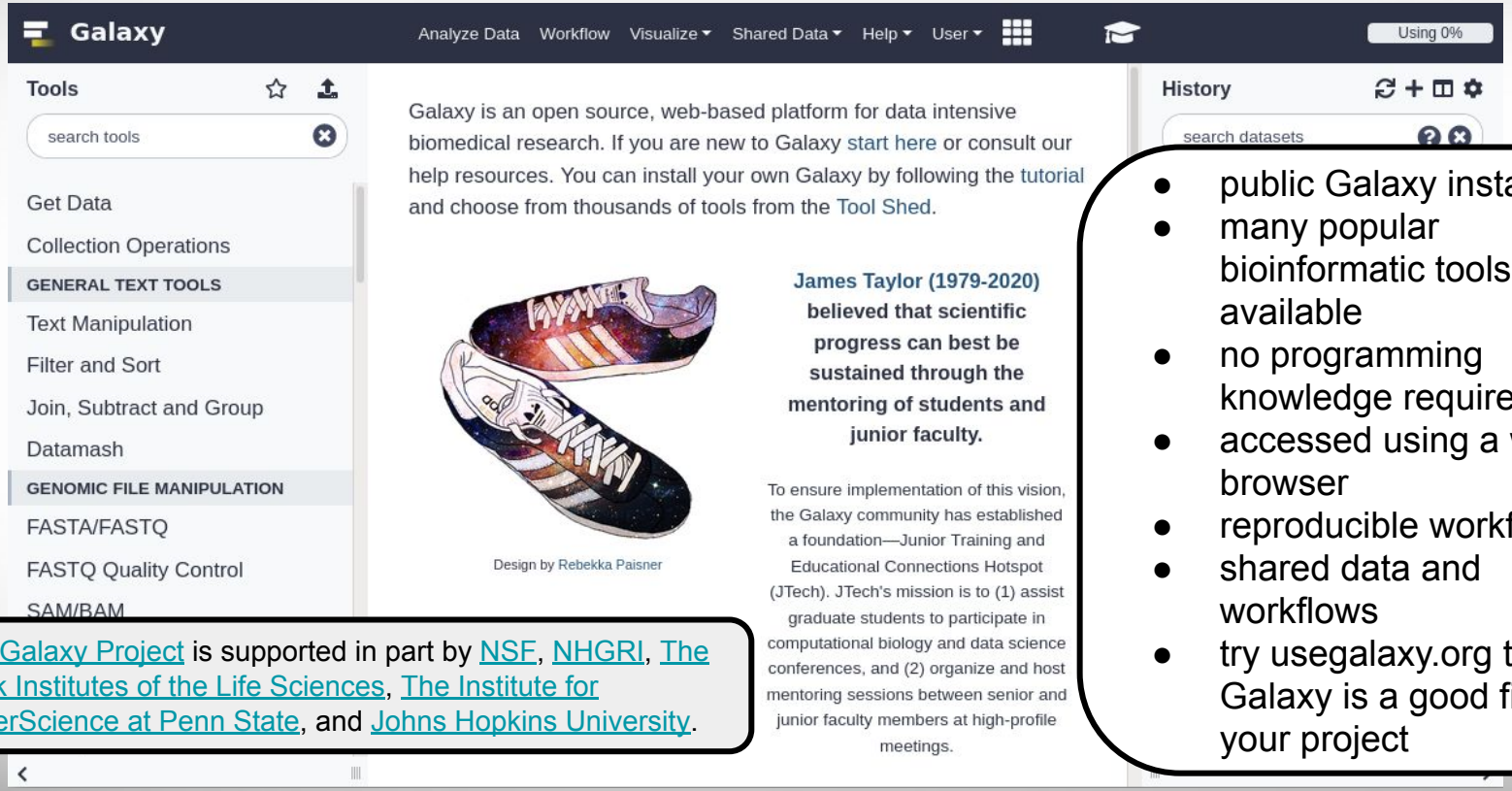
HPRC Maroon Galaxy

What is the Galaxy Project?

usegalaxy.org



- public Galaxy instance
- many popular bioinformatic tools are available
- no programming knowledge required
- accessed using a web browser
- reproducible workflow
- shared data and workflows
- try usegalaxy.org to see if Galaxy is a good fit for your project

The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.

Galaxy Tutorials https://galaxyproject.org/learn

Galaxy COMMUNITY HUB | Use | Learn | Community | Deploy & Develop | Support | Jobs | @jxtx | Search Galaxy | Edit

Learn Galaxy

There are many approaches to learning how to *use* Galaxy. The most popular is probably to just dive in and use it. Galaxy is simple enough to use that you can do many analyses just by exploring the interface. However, you may miss much of the power this way.

Have you created or know of a resource that is useful for teaching with Galaxy? Then please share it! This will help others and also help get the word out about your resource. Use this Google form to describe your resource. **Also**: consider joining Galaxy Training Network and contributing your tutorial as described here!

Tutorials by Galaxy Training Network

Thanks to a large group of wonderful contributors there is a constantly growing set of tutorials maintained by the Galaxy Training Network. These include:

Introductory Tutorials

- Introduction to Galaxy Analyses
- Data Manipulation
- User Interface and Features

Scientific Analyses

- Assembly
- Computational chemistry
- Ecology
- Epigenetics
- Genome Annotation
- Imaging

Tutorials by Galaxy Training Network

Tutorials using Galaxy Main

Interactive Tours

Tutorials from Lewis-Sigler Institute @ Princeton

There are many tutorials available with example input data and step by step analysis for various topics

Galaxy Shared Workflows

usegalaxy.org

- can be used to share your workflow used in a publication
- input files are not stored here

Published Workflows

rna-seq ✕ search name, annotation, owner, and

Advanced Search

Name	Annotation	Owner	Community Rating↓	Community Tags	Last Updated
Basic RNA-Seq Analysis - Differential Expression	From the RNA-Seq analysis tutorial during the Functional Genomics Workshop 2012 https://caps.osu.edu/pfg-workshop	mejia-guerra	★★★★★		Jun 22, 2012
RNA-seq differential analysis (single-end short reads, 2 conditions, 2 replicates)	Workflow based on Tophat and cuffdiff. Inputs: 4 fastq files (experiments), 1 bam file (pseudoreads), 1 gtf file (annotations). Outputs: bam, bigwig, xls,...	rna-seq-helin-group	★★★★★	illumina rnaseq cuffdiff tophat	Jul 17, 2013
RNA-seq preprocessing	Preprocessing of RNA-seq data by quality trimming of read ends and Illumina adapter removal. Inputs: fastq files. Outputs: fastq files and html files with...	rna-seq-helin-group	★★★★★	illumina rnaseq	Jun 03, 2013
RNA-seq differential expression analysis	RNA-seq differential analysis	chmy	★★★★★	easy	Aug 11, 2017

search results for rna-seq and sorted by Community Rating

Galaxy Shared Workflow Example

Galaxy | Workflow | Visualize | Shared Data | Help | User | Using 0%

Tools

search tools

Inputs
Get Data
Collection Operations
Expression Tools
GENERAL TEXT TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
Datamash
GENOMIC FILE MANIPULATION
FASTA/FASTQ
FASTQ Quality Control
SAM/BAM
BED
VCF/BCF
Nanopore
Convert Formats
Lift-Over
COMMON GENOMICS TOOLS

Preproc and TH

Fastq file
output (input)

FASTQ Groomer
File to groom
output_file (fastqsanger, fastqcssanger, fastqsolexa, fastqillumina, fastqsolexa.gz, fastqillumina.gz, fastqsanger.gz, fastqcssanger.gz, fastqsolexa.bz2, fastqillumina.bz2, fastqsanger.bz2, fastqcssanger.bz2)

FASTQ Quality Trimmer
FASTQ File
output_file (input)

Clip
Input file in FASTA or FASTQ format
output (input)

FastQC
Short read data from your current history
Contaminant list
Adapter list
Submodule and Limit specifing file
FastQC on input dataset(s): Webpage (html)
FastQC on input dataset(s): RawData (txt)

FastQC
Short read data from your current history
Contaminant list
Adapter list
Submodule and Limit specifing file
FastQC on input dataset(s): Webpage (html)
FastQC on input dataset(s): RawData (txt)

100%

FastQC Read Quality reports (Galaxy Version 0.72+galaxy1)

Label

Add a step label.

Step Annotation

Add an annotation or notes to this step. Annotations are available when a workflow is viewed.

Short read data from your current history

Data input 'input_file' (fastq, fastq.gz, fastq.bz2, bam or sam)

Contaminant list

Data input 'contaminants' (tabular)

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

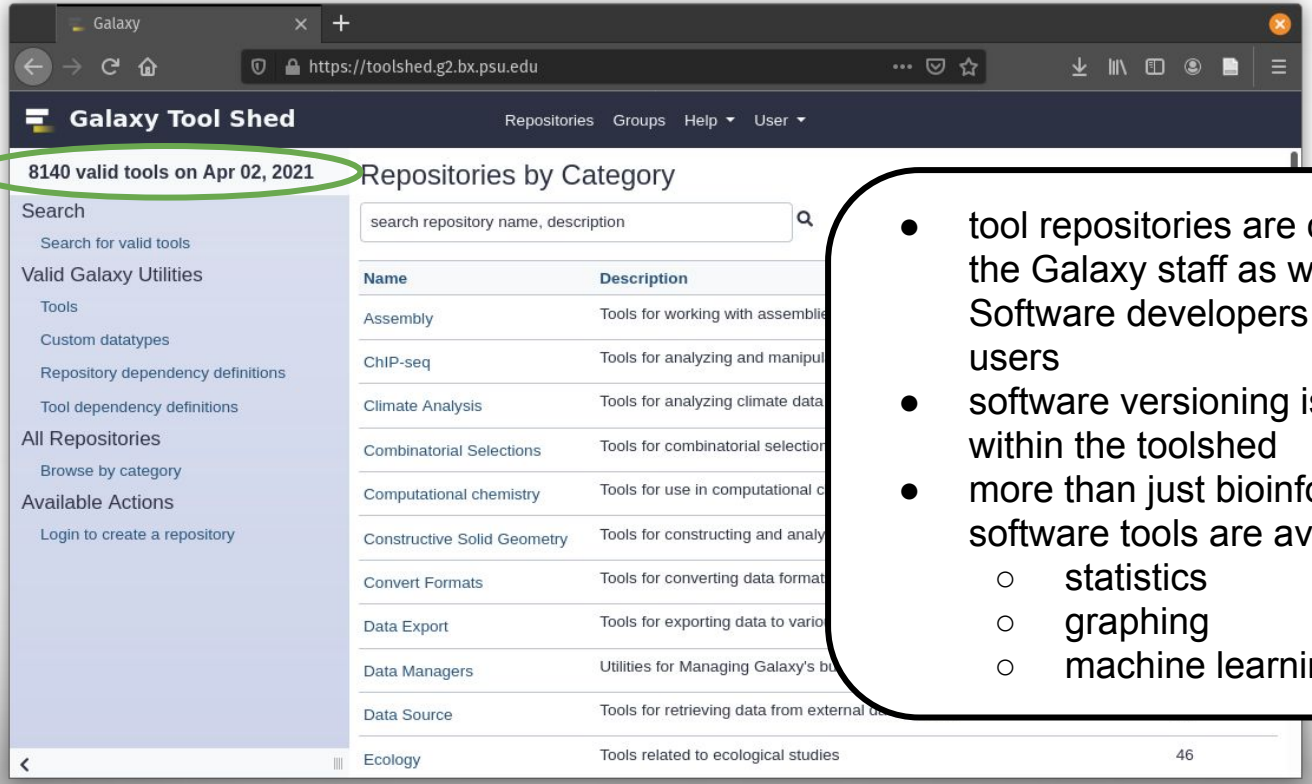
Adapter list

Data input 'adapters' (tabular)

list of adapters adapter sequences

Galaxy Tool Shed

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



- tool repositories are created by the Galaxy staff as well as Software developers and Galaxy users
- software versioning is available within the toolshed
- more than just bioinformatics software tools are available
  - statistics
  - graphing
  - machine learning

The Galaxy Project has Community Support for Feature Requests and Resolving Issues

Issues · galaxyproject/gala

https://github.com/galaxyproject/galaxy/issues

Search or jump to… / Pull requests Issues Marketplace Explore

galaxyproject / galaxy

Watch 71 Star 832 Fork 677

Code Issues 1.5k Pull requests 144 Actions Projects 10 Security Insights

Filters is:issue is:open Labels 72 Milestones 2 New issue

1,468 Open 2,487 Closed Author Label Projects Milestones Assignee Sort

Expand test coverage for new invocation filters + instance parameters area/testing/api help wanted
#11920 opened 7 hours ago by mvdbeek

Deprecate support posgresql < 10 area/database 3
#11919 opened 9 hours ago by mvdbeek 21.05

AttributeError: 'NoneType' object has no attribute 'id' 1
#11917 opened 11 hours ago by maikenp

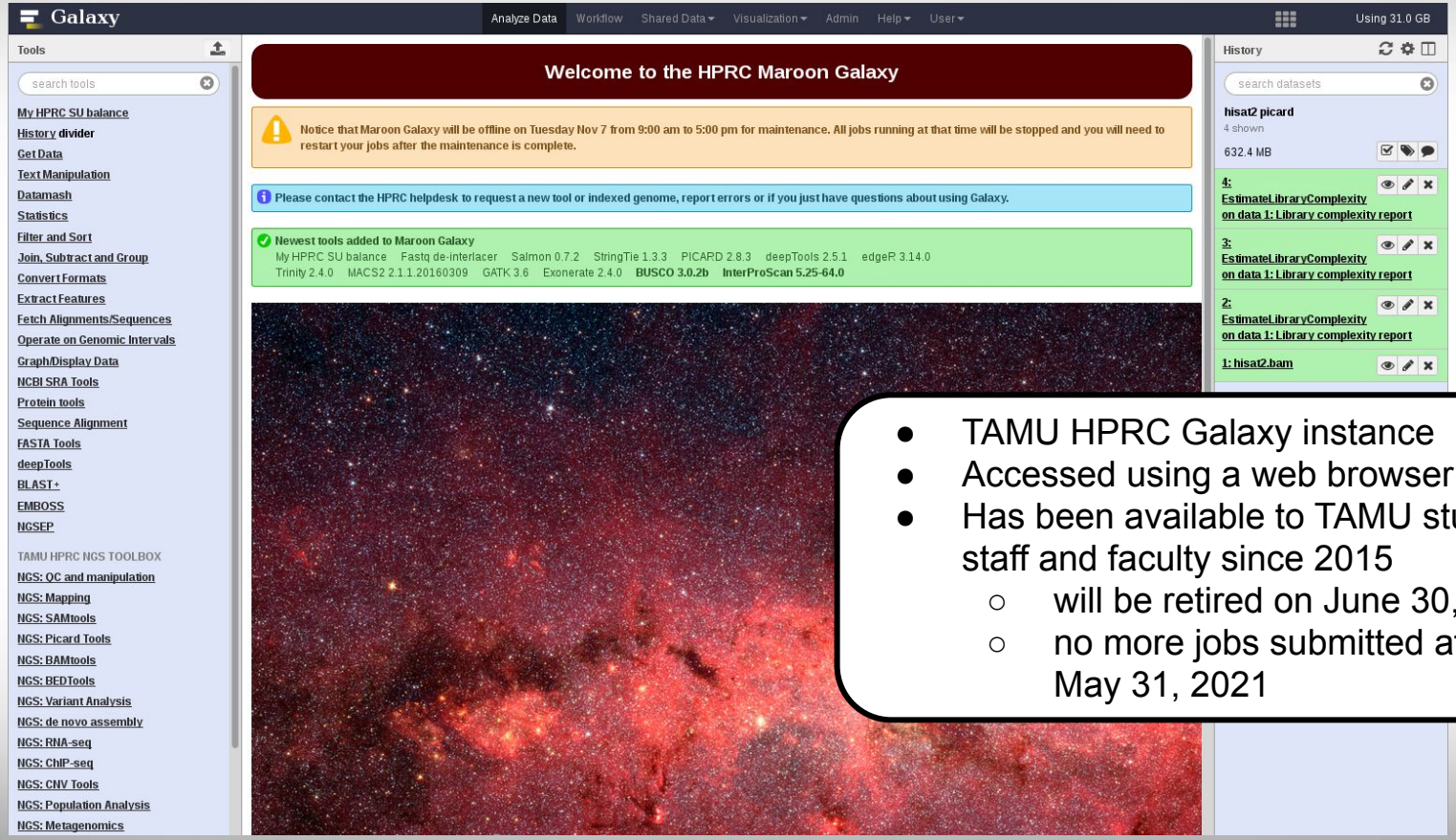
make data libraries respect own filters area/UI-UX area/libraries
#11916 opened 12 hours ago by martenson

uploads hanging (one in a few 1000) 4
#11915 opened 3 days ago by bgruening

Tabular datasets with many columns crash Galaxy on Dataset Peek area/datatypes kind/bug 10
#11885 opened 10 days ago by mtekman

HPRC Maroon Galaxy on the Ada Cluster

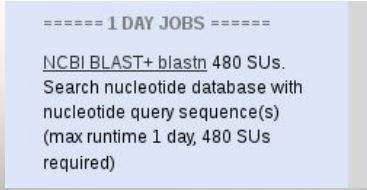
HPRC Maroon Galaxy on the Ada Cluster

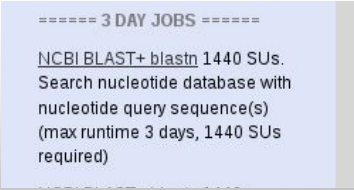


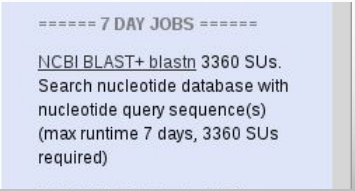
- TAMU HPRC Galaxy instance
- Accessed using a web browser
- Has been available to TAMU students, staff and faculty since 2015
  - will be retired on June 30, 2021
  - no more jobs submitted after May 31, 2021

HPRC Maroon Galaxy on the Ada Cluster

- Installed in 2015 (v15.07) as a joint effort by Ping Luo (HPRC) and Dr. Rodolfo Aramayo (Department of Biology) - (currently 75 registered users)
- There are 350+ users across all HPRC Galaxies that are used for teaching and research
- Auto-installation of tools did not consistently work
- No support for compressed files (gzip) other than decompression on upload
- No auto-detection of fastqsanger format on upload
  - user had to run FastQC to evaluate quality score encoding or run Fastq Groomer
- Most newer software tools require v16.01+
- Early support for running tools on a cluster did not work well so tools were hard coded to use a specific amount of resource (cores, memory, time)
  - multiple tool configurations had to be added to support different resource requirements







HPRC Maroon Galaxy on the Grace Cluster

HPRC Maroon Galaxy on the Grace Cluster

HPRC Maroon Galaxy

Analyze Data | Workflow | Visualize | Shared Data | Admin | Help | User | Using 0%

Tools

search tools

Upload Data

- HPRC
- Get Data
- Send Data
- Collection Operations
- Lift-Over
- Text Manipulation
- Convert Formats
- Filter and Sort
- Join, Subtract and Group
- Fetch Alignments/Sequences
- Operate on Genomic Intervals
- RNA-seq
- Variant Calling

Best Practices for Maroon Galaxy

- Contact the HPRC helpdesk with an email to request a new or updated tool, indexed genome or to report an error. (Maroon Galaxy docs, slides)
- All users begin with a file quota of 1TB. Request an increase if you need more disk space but permanently delete nonessential files before requesting.
- FTP uploads are removed from ftp directory 48 hours after uploading so import your ftp files into Galaxy the same day as you upload to ftp
- The default job resource parameters for all tools is 1 core with 9GB memory for 24 hours (24 SUs).
  - Configuring a job to use 360GB memory for 1 hour requires 48 SUs. (360GB memory for 168 hours = 8,064 SUs).
- Only tools that support multi-core processing have the Job Resources Parameters option which allow you to select cores, memory and time.
  - Configuring a job to use 48 cores and 360GB memory for 1 hour requires 48 SUs. (48 cores for 168 hours = 8,064 SUs).
  - Configuring a job to use 80 cores and 2.93TB memory for 1 hour requires 80 SUs. (80 cores for 168 hours = 13,440 SUs).

COVID-19 related research on Galaxy: training, tutorials, documents

Current known issues:
Some tools cannot be removed from favorites. Choose your favorites wisely.

History

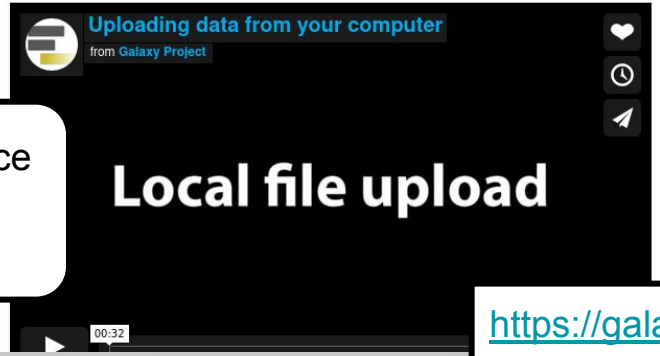
search datasets

Unnamed history

26 shown, 7 deleted

18.47 MB

- 32: _____
- 31: Summary Statistics on data 26
- 30: Summary Statistics on data 26
- 29: test_sam
- 28: Map with minimap2 on data 4 and data 5 (mapped reads in BAM format)
- 26: test.bed
- 23: Map with minimap2 on data 4 and data 5 (mapped reads in BAM format)
- 22: Map with minimap2 on data 4 and data 5 (mapped reads in BAM format)

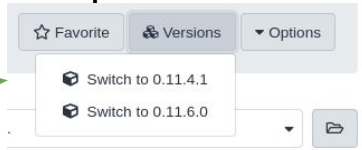


- TAMU HPRC Galaxy instance
- Available now to TAMU students, staff and faculty

https://galaxy-grace.hprc.tamu.edu/maroon

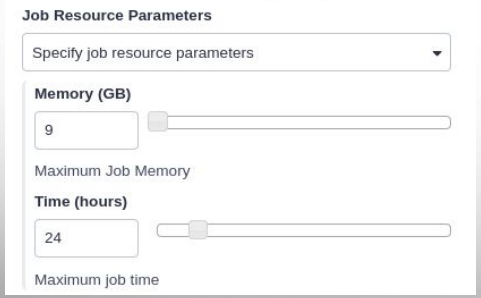
HPRC Grace Maroon Galaxy Features

- Installed in Spring 2021 (v21.01) on the HPRC Grace cluster.
- This is a significant deployment for the HPRC bio community for research and teaching.
- Improved integration of automatically installing tools from the Galaxy Toolshed.
- New support added to maintain file compression after uploading
  - Some software doesn't support compressed file format as input but file decompression is supported
- Multiple tool versions are available within a single tool

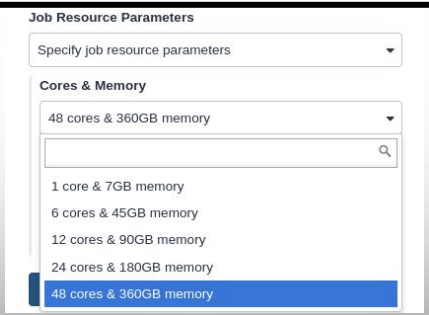


- New auto-detection of fastqsanger format on upload
  - no need to use Fastq Groomer unless working with very early Illumina sequence data
- Integrated support for selecting cluster resources for each tool (cores, memory, time)

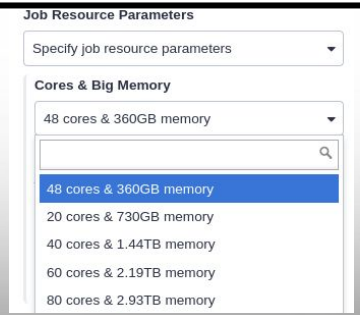
tools supporting **single-core** processing



tools supporting **multi-core** processing



tools supporting **multi-core big memory** processing



Galaxy Tutorials and Chat Features

HPRC Maroon Galaxy Analyze Data Workflow Visualize Shared Data Admin Help User Using 0%

Tools

gffread

Convert Forma

gffread Filters a
GFF3/GTF2 rec

RNA-seq

gffread Filters a
GFF3/GTF2 rec

WORKFLOWS

All workflows

Galaxy Training! Contributors Help Extras Search

Welcome to Galaxy Training!

Collection of tutorials developed and maintained by the worldwide Galaxy community

Galaxy for Scientists

Topic	Tutorials
Introduction to Galaxy Analyses	10
Assembly	9
Climate	4
Computational chemistry	6
Ecology	6

Galaxy Tips & Tricks

Topic	Tutorials
Using Galaxy and Managing your Data	17

Galaxy for Developers and Admins

Topic	Tutorials
Galaxy Server administration	

OPEN CHAT

Click to run unavailable.

Local file upload

The same tutorials found at galaxyproject.org/learn can be accessed here and followed without leaving Maroon Galaxy

More than 300 Galaxy users and admins discussing anything Galaxy related

Circos Tutorial Demonstration

HPRC Maroon Galaxy

Galaxy Training!

Visualisation

Training material for data visualisation in Galaxy

Requirements

Before diving into this topic, we recommend you to have a look at:

- Introduction to Galaxy Analyses

Material

Lesson	Slides	Hands-on	Input dataset	Workflows	Galaxy tour	Galaxy instances
Visualisations in Galaxy						
Genomic Data Visualisation with JBrowse						
Visualisation with Circos						

Check Your HPRC SUs Balance

HPRC Maroon Galaxy | Analyze Data | Workflow | Visualize | Shared Data | Admin | Help | User | Using 0%

Tools

search tools

HPRC

- **My HPRC SU balance**
- **History Divider**

Get Data

Send Data

Collection Operations

Lift-Over

Text Manipulation

Convert Formats

Filter and Sort

Join, Subtract and Group

Fetch Alignments/Sequences

Operate on Genomic Intervals

My HPRC SU balance (Galaxy Version 1.0.0) | Favorite | Options

I want to

Show my current SU balance

Execute

If you are unable to run this tool then you either need to renew your HPRC account or you are out of SUs.

This tool retreives a summary of your HPRC SU balance or allows the user to set the default account.

Run this tool selecting the option 'Show my current SU balance' to get a list of your project account numbers.

Account	Default	Allocation	Used & Pending SUs	Balance
000000000001	Y	5000.00	-4990.00	10.00

History

search datasets

trinity

45 shown, 110 deleted, 2 hidden

7.4 GB

- **154: My HPRC SU balance**
- **153: Interproscan calculation on data 104**
- **152: Pasted Entry**
- **151: Interproscan calculation on data 103**
- **150: Interproscan calculation on data 103**
- **149: Interproscan calculation on data 103**
- **147: Interproscan calcul**

You can also change your default HPRC project account and view SUs charged for each job run during the current fiscal year

Shared Data Libraries

HPRC Maroon Galaxy | Analyze Data | Workflow | Visualize | Shared Data | Admin | Help | User | Using 0%

Search | + Folder | + Datasets | Export to His | Data Libraries | Histories | Workflows | Visualizations | Pages | Delete | Details | include deleted

Libraries / C_dubliniensis_CD36 data

Name	Description	Da	Date Updated (UTC)	State
..				
DR34_R1.fastq.gz	uploaded fastqsanger.gz file	fastqsanger.gz 162.4 MB	2020-11-09 08:28 PM	Manage

« 0 1 2 » 15 per page, 1 total

Files can be added to a 'Data Library' which you can share with your Galaxy colleagues. Send a request to the HPRC helpdesk if you would like a Data Library for your group

NCBI Blast Databases Available

NCBI BLAST+ blastn Search nucleotide database with nucleotide query sequence(s) (Galaxy Version 2.10.1+galaxy0)

Favorite | Options

Nucleotide query sequence(s)

3: M.tuberculosis_genome.fasta

(-query)

Subject database/sequences

Locally installed BLAST database

Nucleotide BLAST database

Select/Unselect all

× NCBI nt (partially non-redundant) 12 Apr 2021

Type of BLAST

megablast - Traditional megablast used to find very similar (e.g., intraspecies or closely related species) sequences

blastn - Traditional BLASTN requiring an exact match of 11, f

blastn-short - BLASTN program optimized for sequences sho

dc-megablast - Discontiguous megablast used to find more di

Additional NCBI BLAST and custom databases can be added upon request

HPRC Maroon Galaxy Access

- Try Galaxy at usegalaxy.org to see if it appropriate for your project
- How to access HPRC Maroon Galaxy on the Grace Cluster
  - Available to Texas A&M students, staff and faculty with a NetID and an HPRC account
  - Apply for an HPRC account first
    - https://hprc.tamu.edu/apply
  - Then send an email request for a Maroon Galaxy account
    - help@hprc.tamu.edu
  - Need to use TAMU VPN when connecting to Galaxy from off campus
  - Login to Maroon Galaxy using your TAMU NetID and password
- Read the Galaxy Usage Notes
  - https://hprc.tamu.edu/wiki/SW:Galaxy

HIGH PERFORMANCE RESEARCH COMPUTING
TEXAS A&M UNIVERSITY

Thank you.

Any questions?