Galaxy as an educational tool and community resources for undergraduate training

PAG 2020
Slides: http://bit.do/teach-gxy-pag-2020

Mo Heydarian
Johns Hopkins University

Dave Clements
Johns Hopkins University

Galaxy Team
https://galaxyproject.org
https://help.galaxyproject.org

#usegalaxy @galaxyproject

Goals for this session

Provide an *introduction* to using Galaxy for bioinformatic analysis.

Demonstrate features of Galaxy that promote accessibility, reproducibility, and transparency.

Highlight Galaxy components and capabilities to leverage for teaching.

Recommendations on resource usage.

What is Galaxy?

Galaxy is an open-source web-based framework engineered to handle large data reproducibly and transparently.

What is Galaxy?

Galaxy is an open-source web-based framework engineered to handle large data reproducibly and transparently.

Users interact with data and tools via a graphical user interface. No computational experience required.

What is Galaxy?

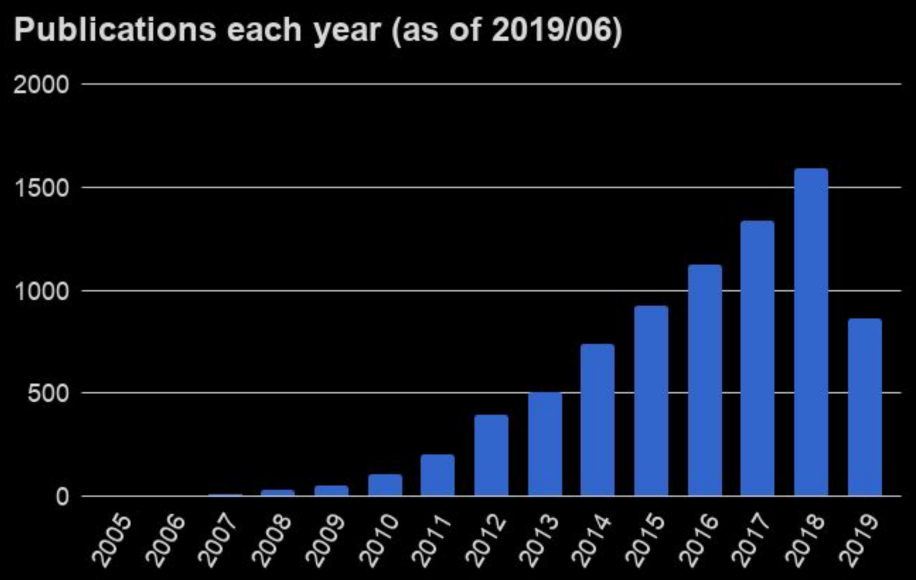
Galaxy is an open-source web-based framework engineered to handle large data reproducibly and transparently.

Users interact with data and tools via a graphical user interface. No computational experience required.

Students do not require coding experience or understanding to use Galaxy.

Who uses Galaxy?

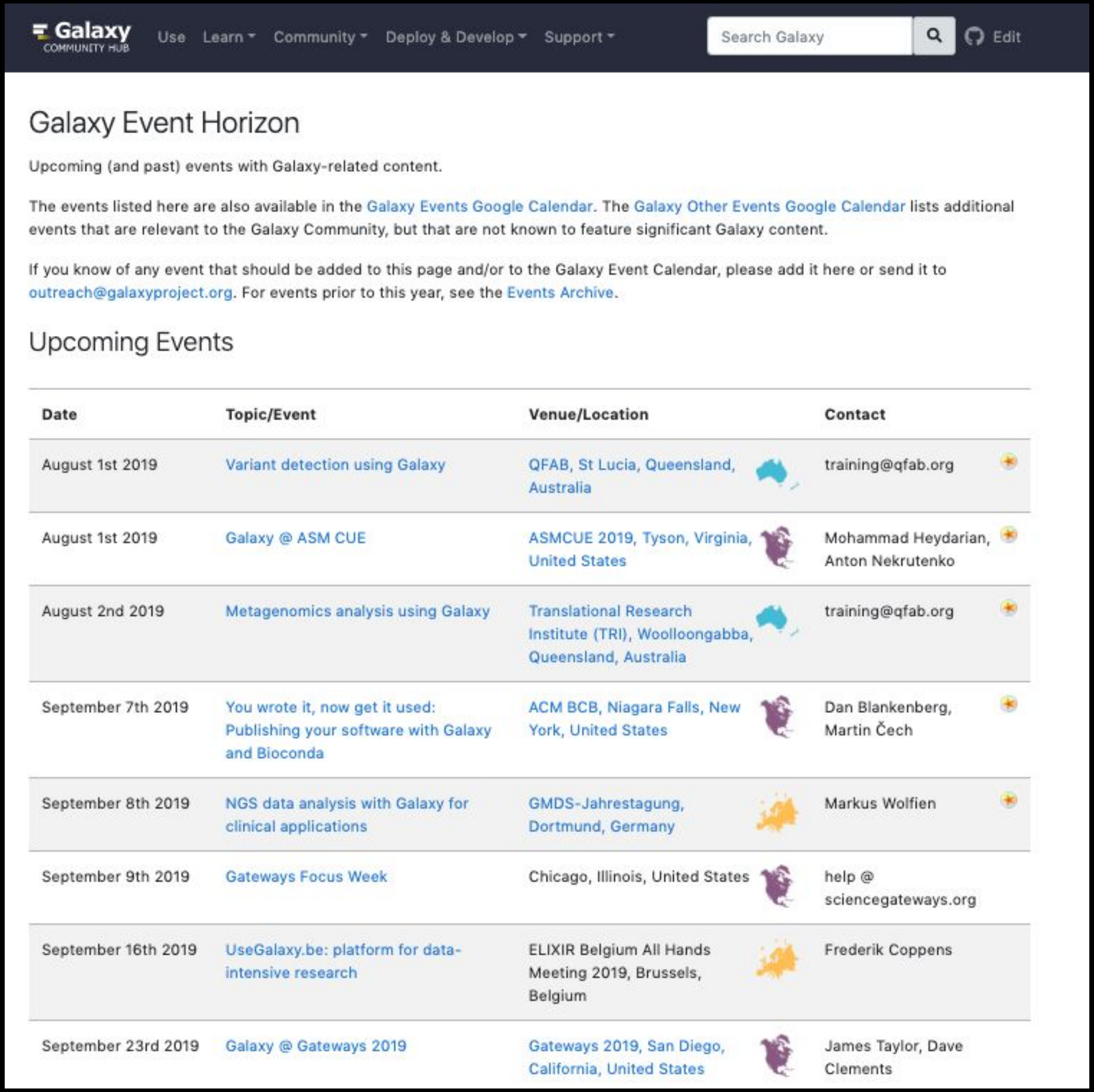
- Scientific researchers across diverse domains
  - Genomics, proteomics, metabolomics, computational chemistry, ecology, natural language processing, climate science, image processing, immunology, single cell analysis, and expanding!
  - Six continents
  - Academia, pharma, government agencies.
  - Galaxy cited in 8,000 publications


Publications each year (as of 2019/06)

https://galaxyproject.org/galaxy-project/statistics/

Who uses Galaxy?

- Teachers!
  - Galaxy is a great teaching platform.
    - GUI access
    - Good support
    - Great community
  - Over 100 Galaxy training events per year.



Galaxy Event Horizon

Upcoming (and past) events with Galaxy-related content.

The events listed here are also available in the Galaxy Events Google Calendar. The Galaxy Other Events Google Calendar lists additional events that are relevant to the Galaxy Community, but that are not known to feature significant Galaxy content.

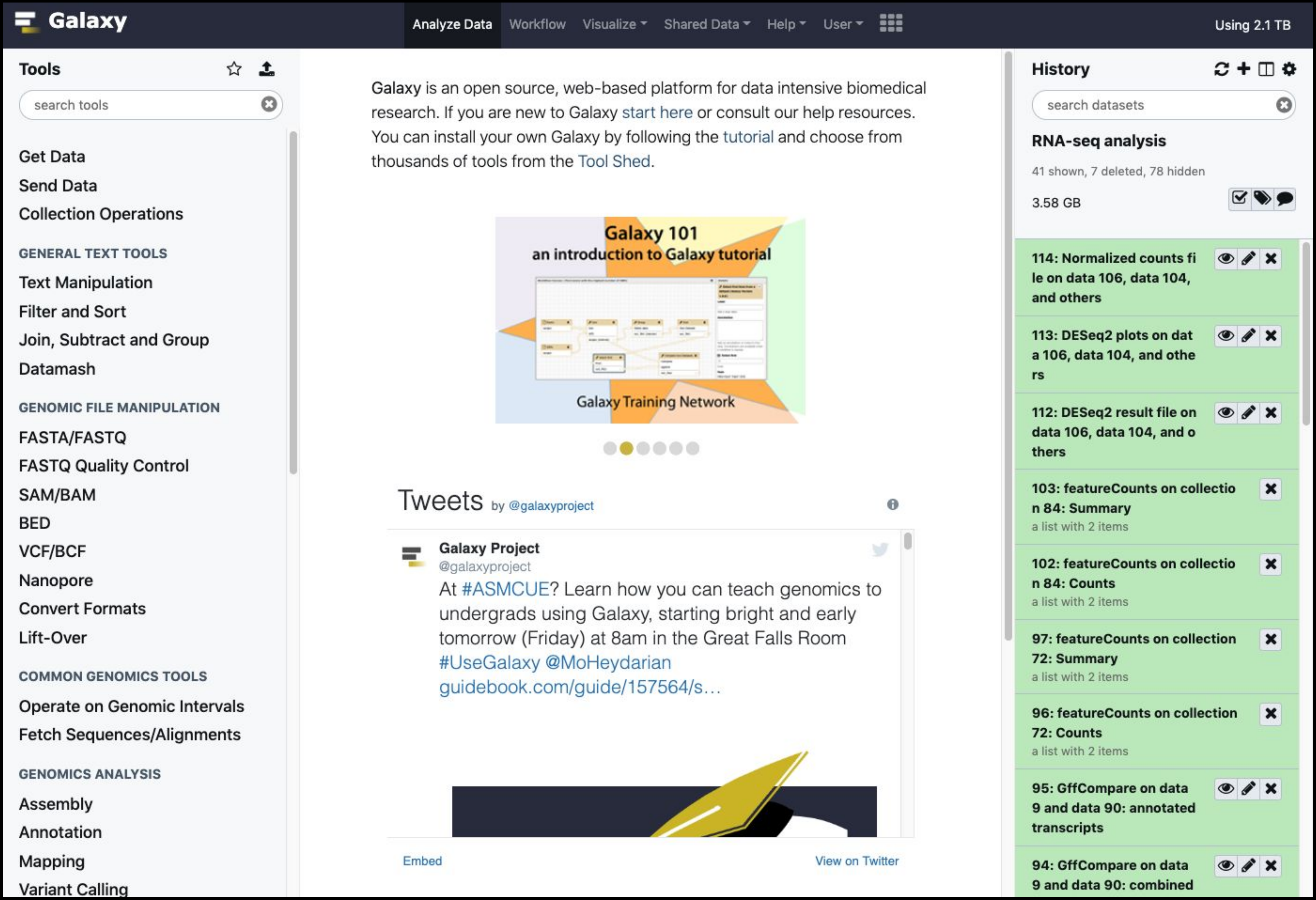
If you know of any event that should be added to this page and/or to the Galaxy Event Calendar, please add it here or send it to outreach@galaxyproject.org. For events prior to this year, see the Events Archive.

Upcoming Events

Date	Topic/Event	Venue/Location	Contact
August 1st 2019	Variant detection using Galaxy	QFAB, St Lucia, Queensland, Australia	training@qfab.org
August 1st 2019	Galaxy @ ASM CUE	ASMCUE 2019, Tyson, Virginia, United States	Mohammad Heydarian, Anton Nekrutenko
August 2nd 2019	Metagenomics analysis using Galaxy	Translational Research Institute (TRI), Woolloongabba, Queensland, Australia	training@qfab.org
September 7th 2019	You wrote it, now get it used: Publishing your software with Galaxy and Bioconda	ACM BCB, Niagara Falls, New York, United States	Dan Blankenberg, Martin Čech
September 8th 2019	NGS data analysis with Galaxy for clinical applications	GMDS-Jahrestagung, Dortmund, Germany	Markus Wolfien
September 9th 2019	Gateways Focus Week	Chicago, Illinois, United States	help @ sciencegateways.org
September 16th 2019	UseGalaxy.be: platform for data-intensive research	ELIXIR Belgium All Hands Meeting 2019, Brussels, Belgium	Frederik Coppens
September 23rd 2019	Galaxy @ Gateways 2019	Gateways 2019, San Diego, California, United States	James Taylor, Dave Clements

https://galaxyproject.org/events/

The Galaxy interface



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. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

Galaxy Project
@galaxyproject
At #ASMCUE? Learn how you can teach genomics to undergrads using Galaxy, starting bright and early tomorrow (Friday) at 8am in the Great Falls Room #UseGalaxy @MoHeydarian guidebook.com/guide/157564/s…

RNA-seq analysis
41 shown, 7 deleted, 78 hidden
3.58 GB

114: Normalized counts file on data 106, data 104, and others

113: DESeq2 plots on data 106, data 104, and others

112: DESeq2 result file on data 106, data 104, and others

103: featureCounts on collection 84: Summary
a list with 2 items

102: featureCounts on collection 84: Counts
a list with 2 items

97: featureCounts on collection 72: Summary
a list with 2 items

96: featureCounts on collection 72: Counts
a list with 2 items

95: GffCompare on data 9 and data 90: annotated transcripts

94: GffCompare on data 9 and data 90: combined

Some Galaxy Terminology

Dataset

Any input, output, or intermediate dataset and any associated metadata

History

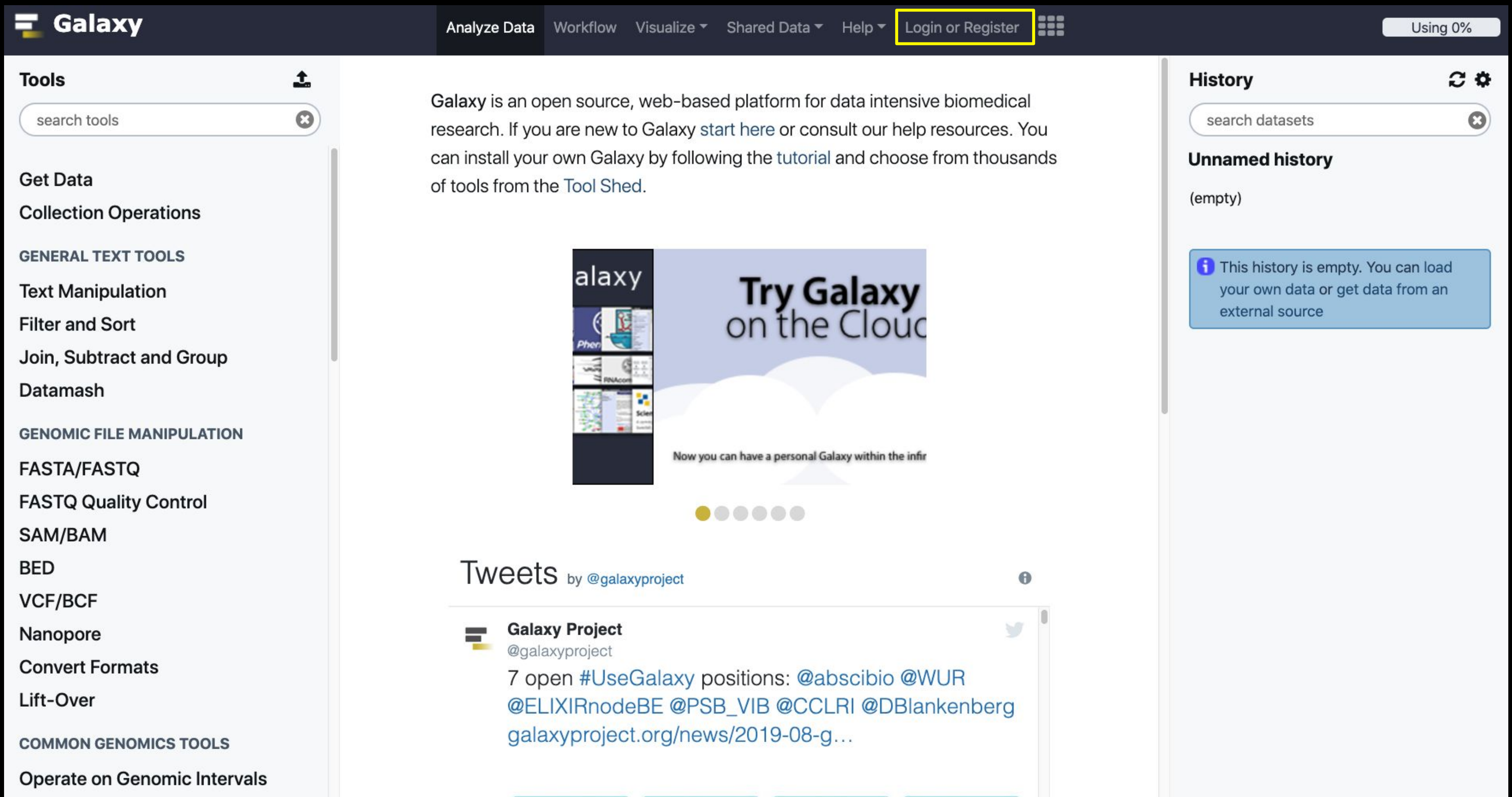
A record of inputs, analysis steps, intermediate datasets, and outputs

Workflow

A series of analysis steps which can be repeated with different data

Hands on exploration of Galaxy

- Go to usegalaxy.org
  - Login or register an account



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. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

This history is empty. You can load your own data or get data from an external source

Galaxy Project
@galaxyproject
7 open #UseGalaxy positions: @abscibio @WUR @ELIXIRnodeBE @PSB_VIB @CCLRI @DBlankenberg galaxyproject.org/news/2019-08-g…

Data upload

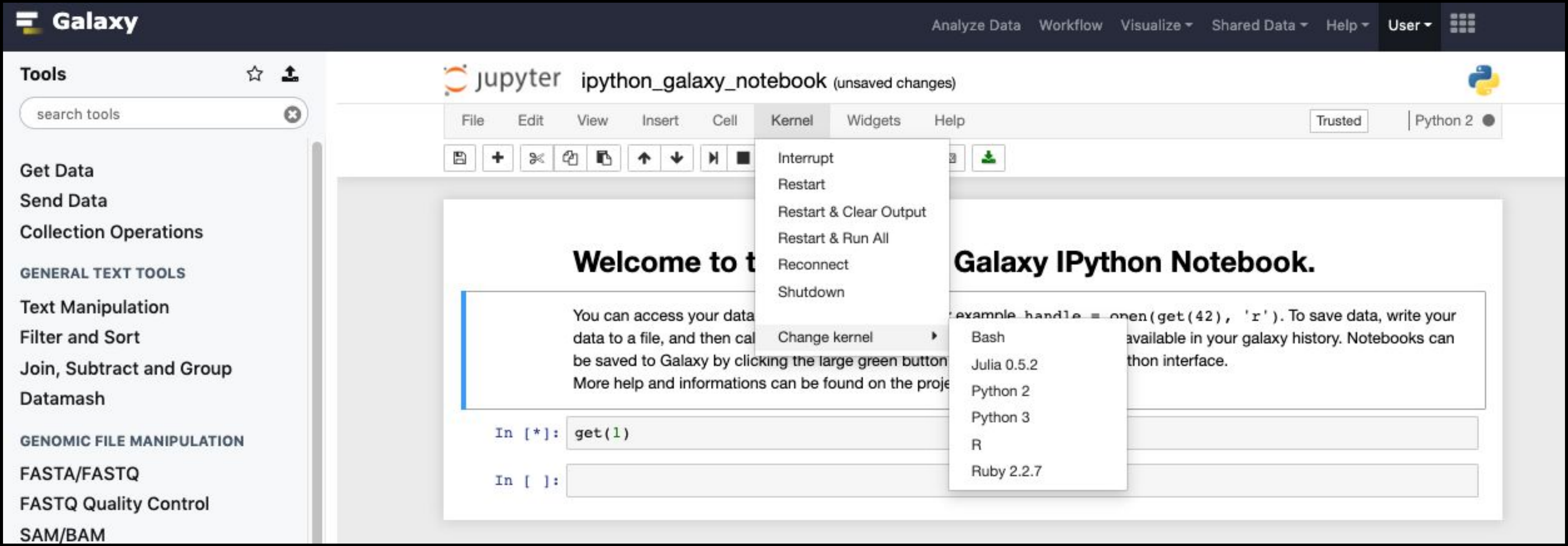
- Upload from interface
- Import from external sources
  - With SRA tools and ENA
  - UCSC table browser
  - Data libraries

Tools

- Organized by categories and searchable
- Broad range
  - Simple UNIX operations
  - NGS read QC, alignment, quantification
  - Visualization
- Tool form with ‘help’ and adjustable parameters in center pane
- Analysis capabilities extended with Galaxy Interactive Environments

Galaxy interactive environments

- Jupyter notebook extends analysis capabilities within Galaxy
  - Kernels for python 2/3, R, Bash, Ruby, and Julia



Welcome to t **Galaxy IPython Notebook.**

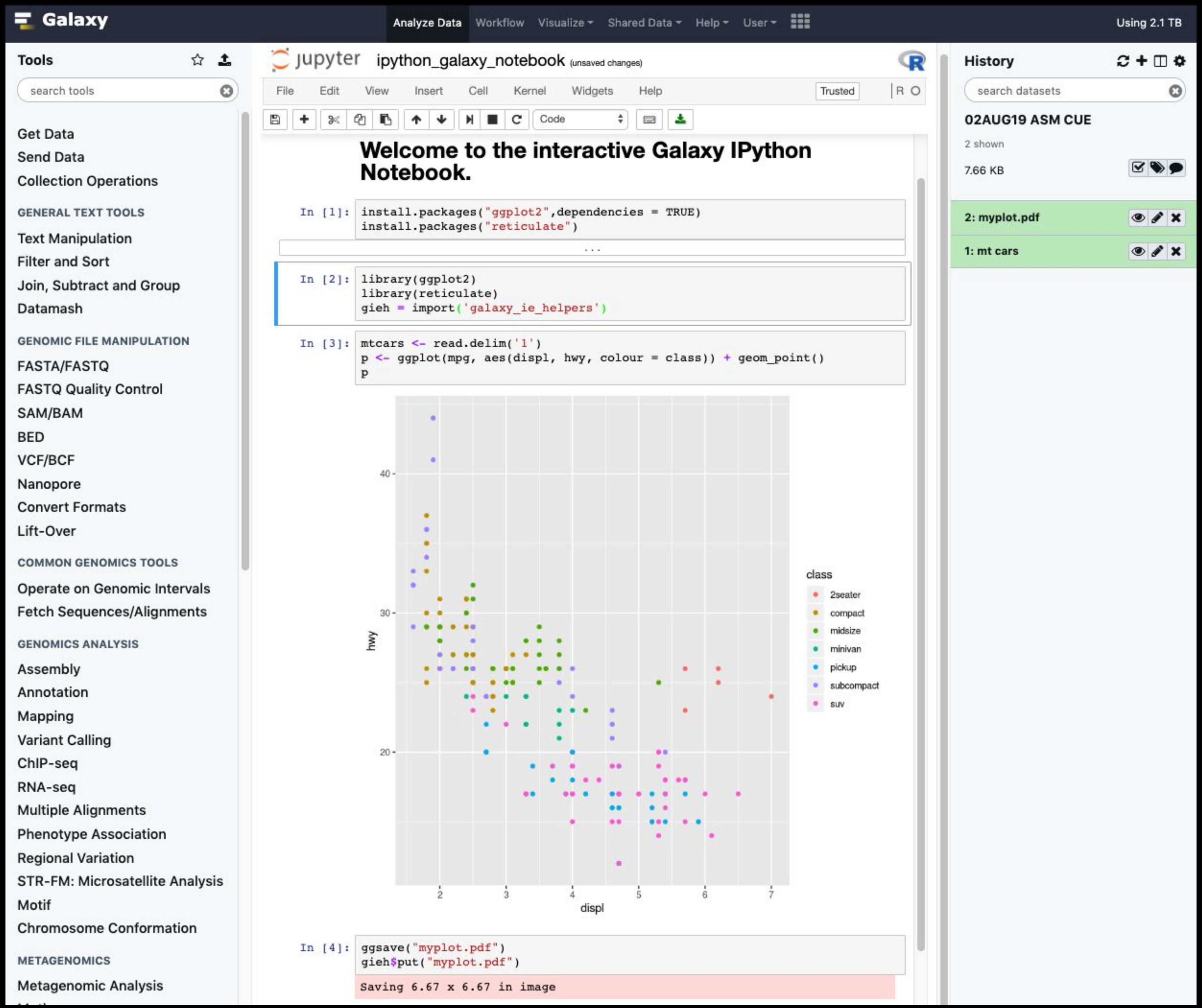
You can access your data example `handle = open(get(42), 'r')`. To save data, write your data to a file, and then cal available in your galaxy history. Notebooks can be saved to Galaxy by clicking the large green button thon interface. More help and informations can be found on the proje

```
In [*]: get(1)
```

```
In [ ]:
```

Galaxy interactive environments (GIEs)

- Jupyter notebook extends analysis capabilities within Galaxy
  - Accessible to importing additional packages
  - Export data items back to history
  - Save and reload notebook, or download
- GIE Tutorial and usage



Welcome to the interactive Galaxy IPython Notebook.

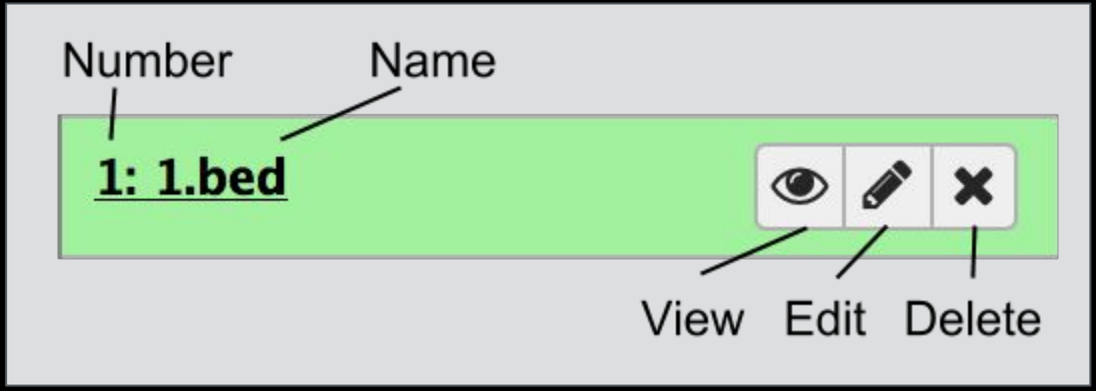
```
In [1]: install.packages("ggplot2",dependencies = TRUE)
        install.packages("reticulate")

In [2]: library(ggplot2)
        library(reticulate)
        gieh = import('galaxy_ie_helpers')

In [3]: mtcars <- read.delim('1')
        p <- ggplot(mpg, aes(displ, hwy, colour = class)) + geom_point()
        p

In [4]: ggsave("myplot.pdf")
        gieh$put("myplot.pdf")

Saving 6.67 x 6.67 in image
```

^ Notebook run on usegalaxy.org

History and data management

- Anatomy of a dataset
- History information and attributes
- Operating on multiple data items
  - Collections and building complex collections with Rule Builder
- Data set state
- Deleting datasets (or not)
- History menu
  - Share, copy, extract workflows

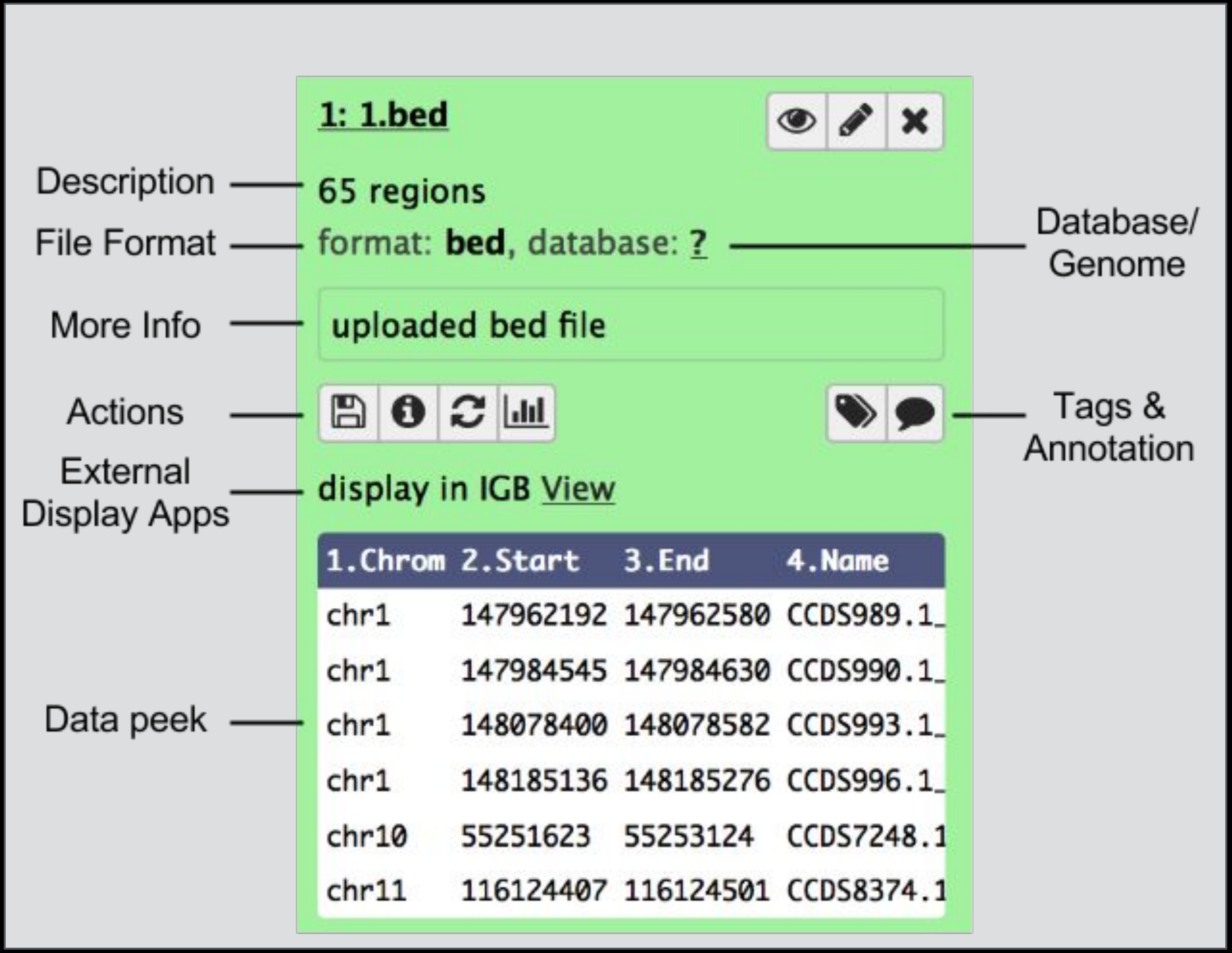
History and data management



- **Anatomy of a dataset**
- History information and attributes
- Operating on multiple data items
  - Collections and building complex collections with Rule Builder
- Data set state
- Deleting datasets (or not)

History and data management

- **Anatomy of a dataset**
- History information and attributes
- Operating on multiple data items
  - Collections and building complex collections with Rule Builder
- Data set state
- Deleting datasets (or not)



1: 1.bed

65 regions

format: **bed**, database: ?

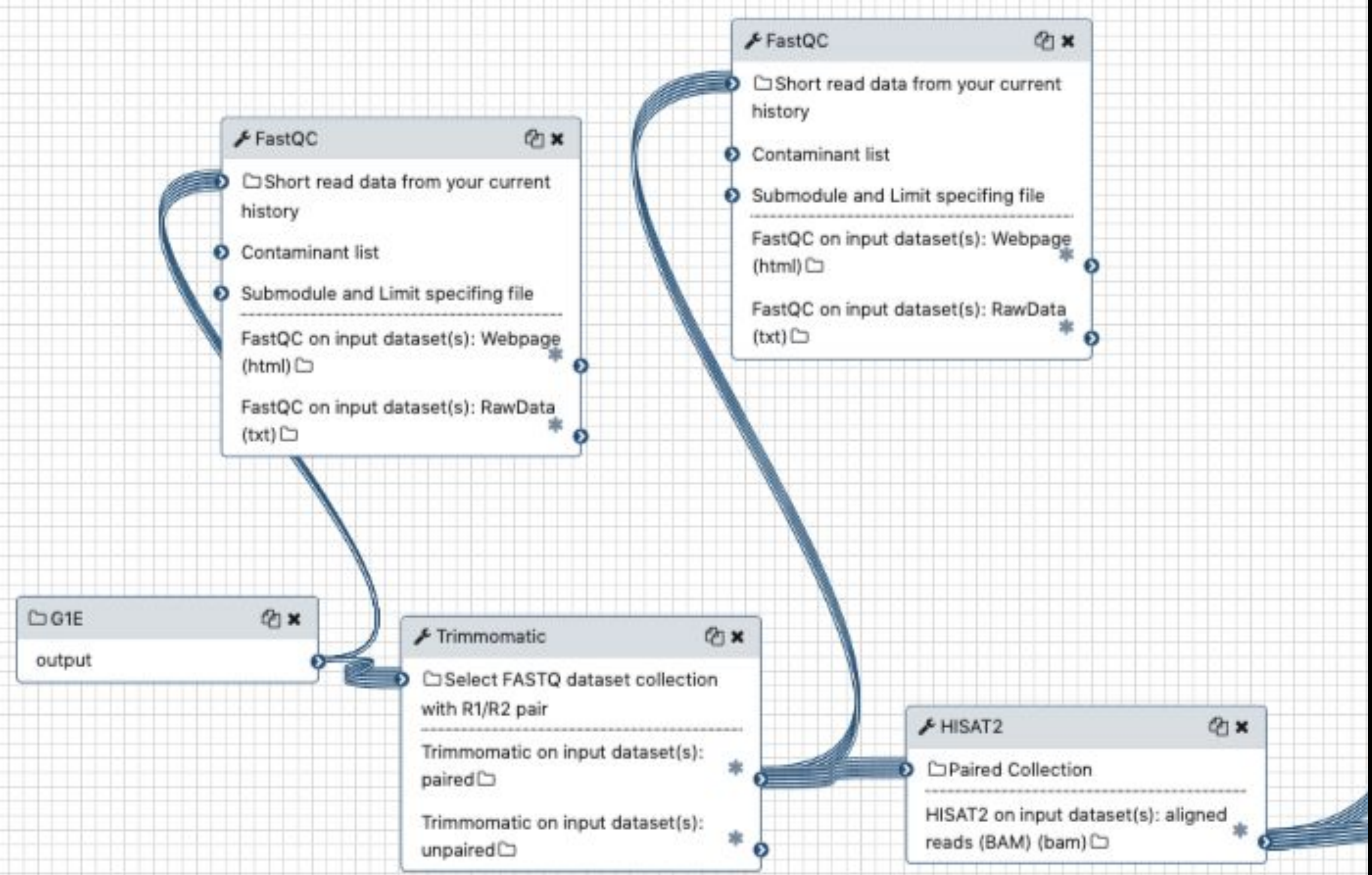
uploaded bed file

display in IGB View

1.Chrom	2.Start	3.End	4.Name
chr1	147962192	147962580	CCDS989.1_
chr1	147984545	147984630	CCDS990.1_
chr1	148078400	148078582	CCDS993.1_
chr1	148185136	148185276	CCDS996.1_
chr10	55251623	55253124	CCDS7248.1
chr11	116124407	116124501	CCDS8374.1

Reproducible analysis with Workflows

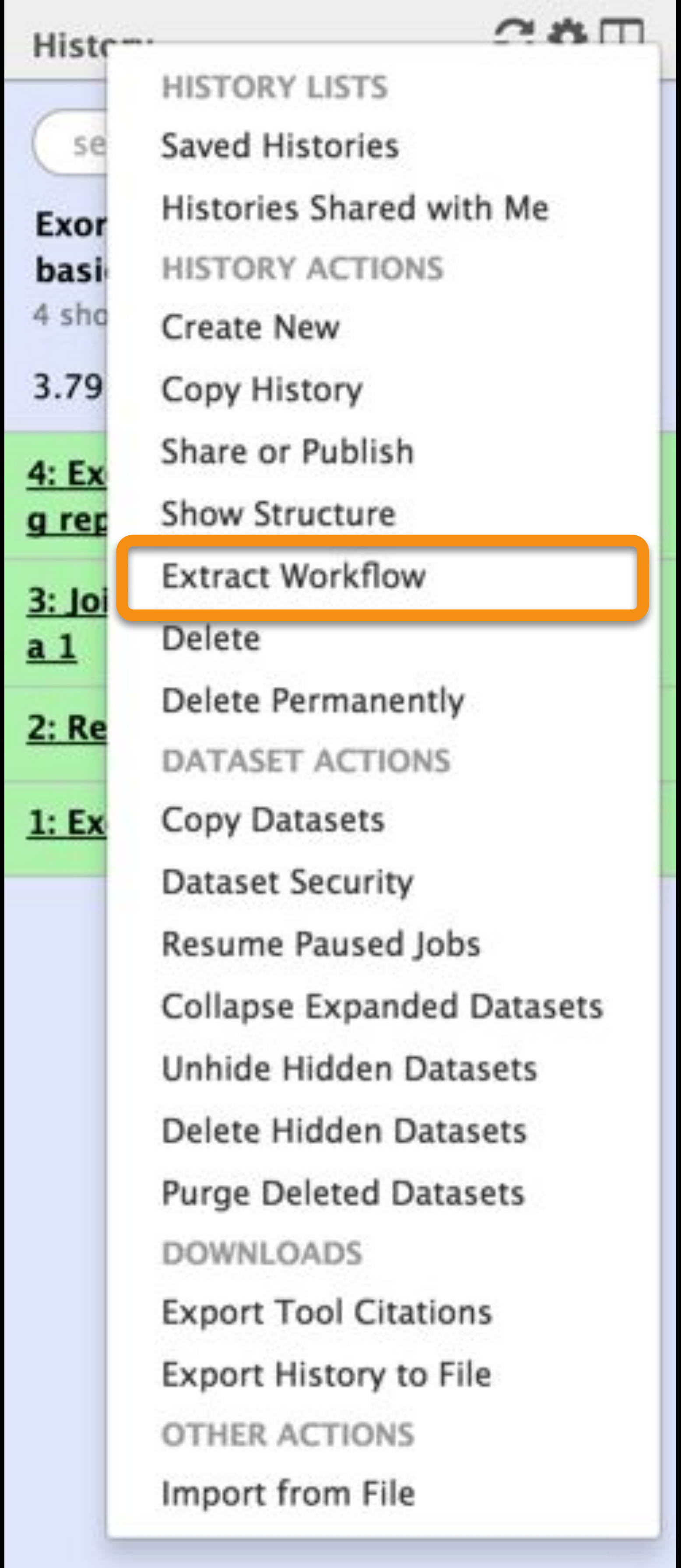
- Chain tools together to create executable analysis pipelines
- Modify tool parameters, change data types, and rename datasets for iterative analysis
- Extract workflows from existing analysis steps

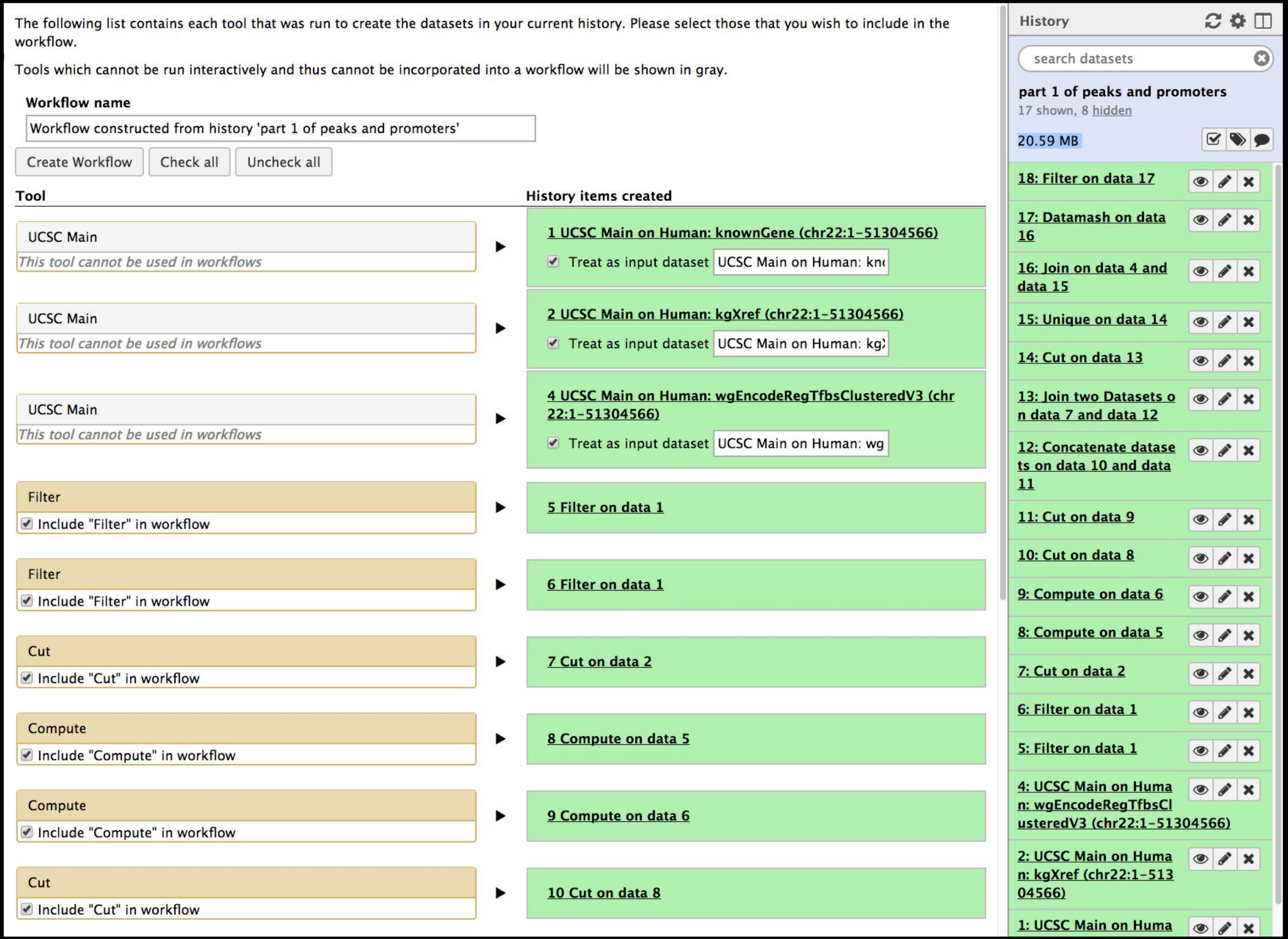


Extract Workflow from history

Create a workflow from this history

⚙ (cog) → Extract Workflow





The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow.

Tools which cannot be run interactively and thus cannot be incorporated into a workflow will be shown in gray.

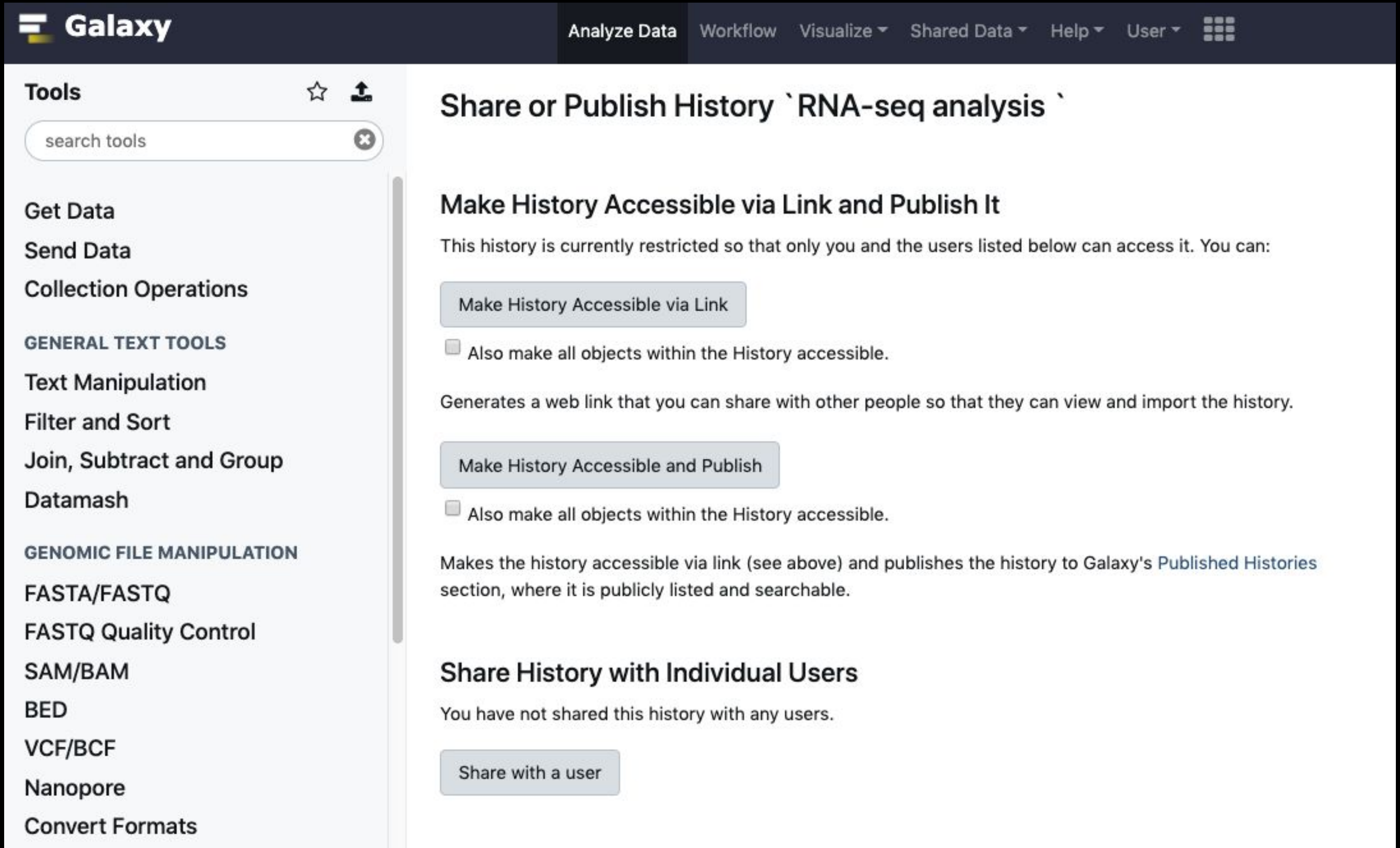
Workflow name

Workflow constructed from history 'part 1 of peaks and promoters'

Tool	History items created
UCSC Main *This tool cannot be used in workflows*	1 UCSC Main on Human: knownGene (chr22:1-51304566) ☑ Treat as input dataset UCSC Main on Human: knc
UCSC Main *This tool cannot be used in workflows*	2 UCSC Main on Human: kgXref (chr22:1-51304566) ☑ Treat as input dataset UCSC Main on Human: kg)
UCSC Main *This tool cannot be used in workflows*	4 UCSC Main on Human: wgEncodeRegTfbsClusteredV3 (chr22:1-51304566) ☑ Treat as input dataset UCSC Main on Human: wg
Filter ☑ Include "Filter" in workflow	5 Filter on data 1
Filter ☑ Include "Filter" in workflow	6 Filter on data 1
Cut ☑ Include "Cut" in workflow	7 Cut on data 2
Compute ☑ Include "Compute" in workflow	8 Compute on data 5
Compute ☑ Include "Compute" in workflow	9 Compute on data 6
Cut ☑ Include "Cut" in workflow	10 Cut on data 8

Accessibility via Sharing features

- Share histories and workflows to defined users
- Publish histories and workflows to all users of a Galaxy instance



Share or Publish History `RNA-seq analysis `

Make History Accessible via Link and Publish It

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

☐ Also make all objects within the History accessible.

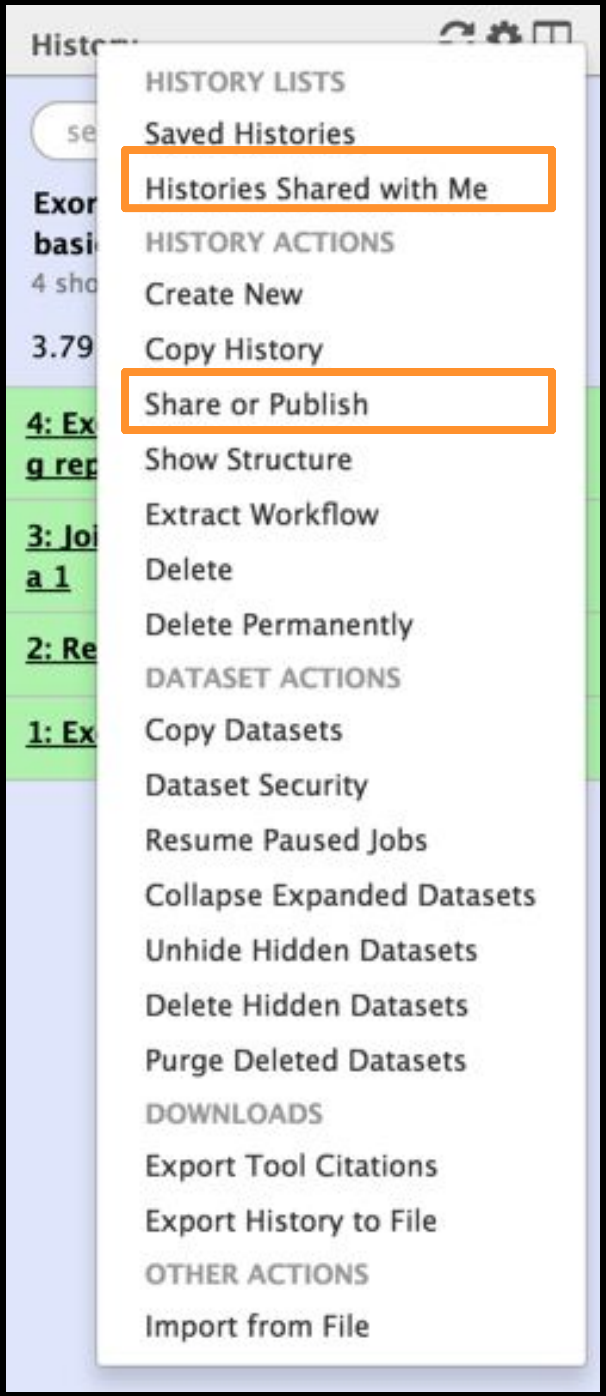
Generates a web link that you can share with other people so that they can view and import the history.

☐ Also make all objects within the History accessible.

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.

Share History with Individual Users

You have not shared this history with any users.

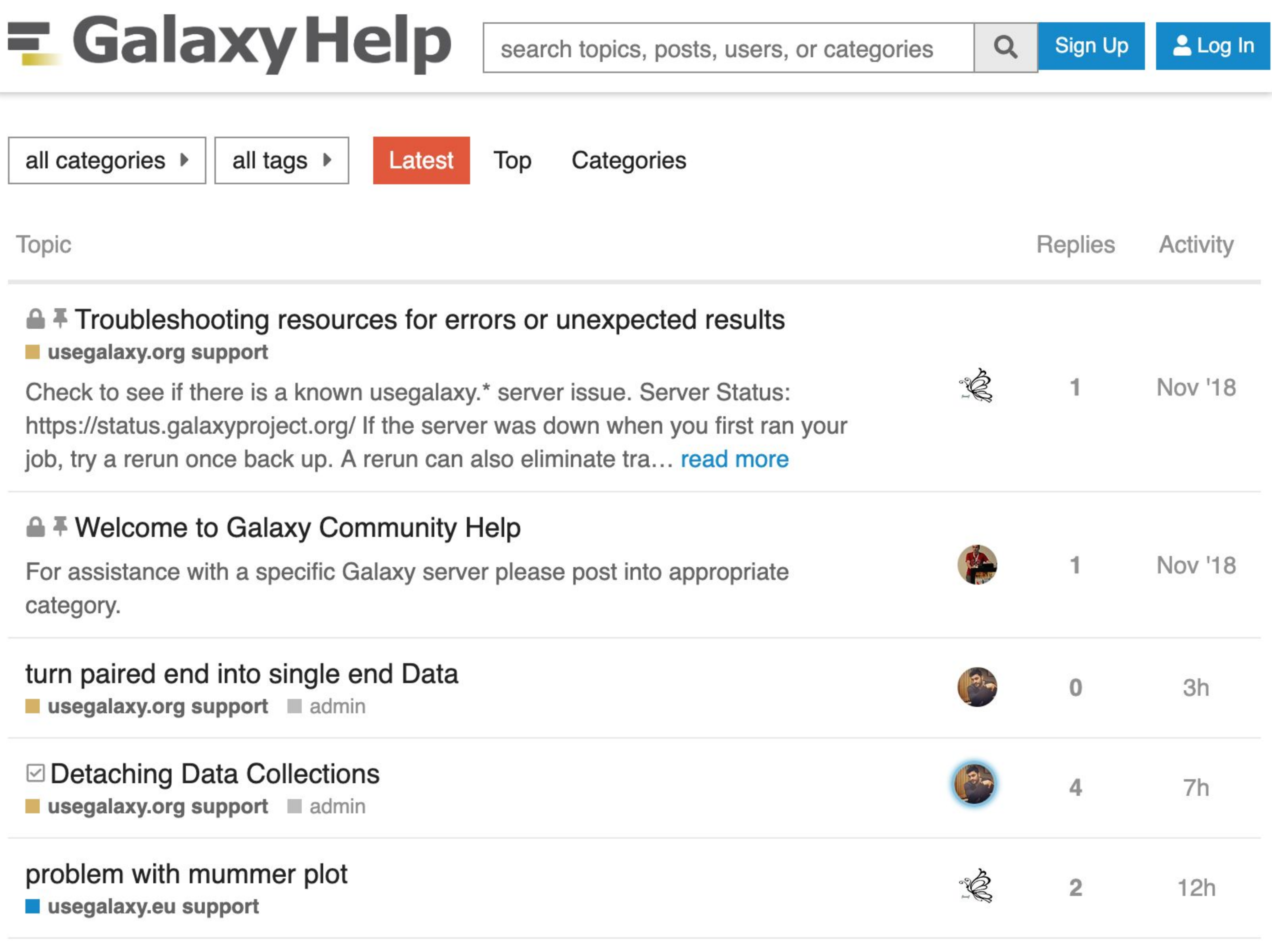


Accessibility via Sharing features

- Data libraries can be populated with shareable data, but requires admin privileges
    - All users have access to data libraries
    - All Galaxy Training Material sample data is available on Data Libraries of useGalaxy.* servers

Learning and support

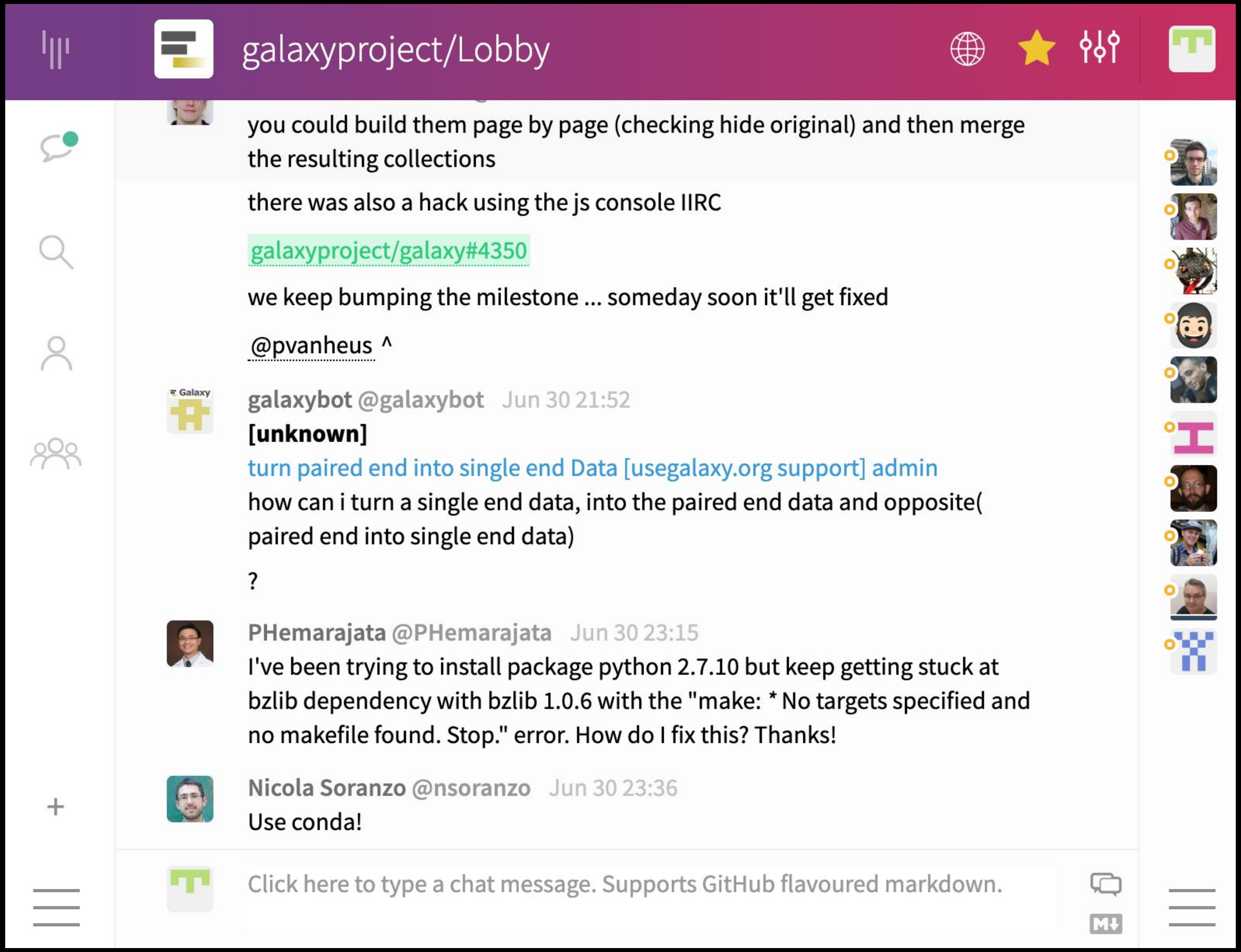
- Support
  - **Help.GalaxyProject.org**
  - Gitter
  - Direct reporting



Topic	Replies	Activity
Troubleshooting resources for errors or unexpected results usegalaxy.org support Check to see if there is a known usegalaxy.* server issue. Server Status: https://status.galaxyproject.org/ If the server was down when you first ran your job, try a rerun once back up. A rerun can also eliminate tra… read more	1	Nov '18
Welcome to Galaxy Community Help For assistance with a specific Galaxy server please post into appropriate category.	1	Nov '18
turn paired end into single end Data usegalaxy.org support admin	0	3h
☑ Detaching Data Collections usegalaxy.org support admin	4	7h
problem with mummer plot usegalaxy.eu support	2	12h

Learning and support

- Support
  - Help.GalaxyProject.org
  - **Gitter**
  - Direct reporting



you could build them page by page (checking hide original) and then merge the resulting collections

there was also a hack using the js console IIRC

galaxyproject/galaxy#4350

we keep bumping the milestone ... someday soon it'll get fixed

@pvanheus ^

galaxybot @galaxybot Jun 30 21:52

[unknown]

turn paired end into single end Data [usegalaxy.org support] admin

how can i turn a single end data, into the paired end data and opposite(paired end into single end data)

?

PHemarajata @PHemarajata Jun 30 23:15

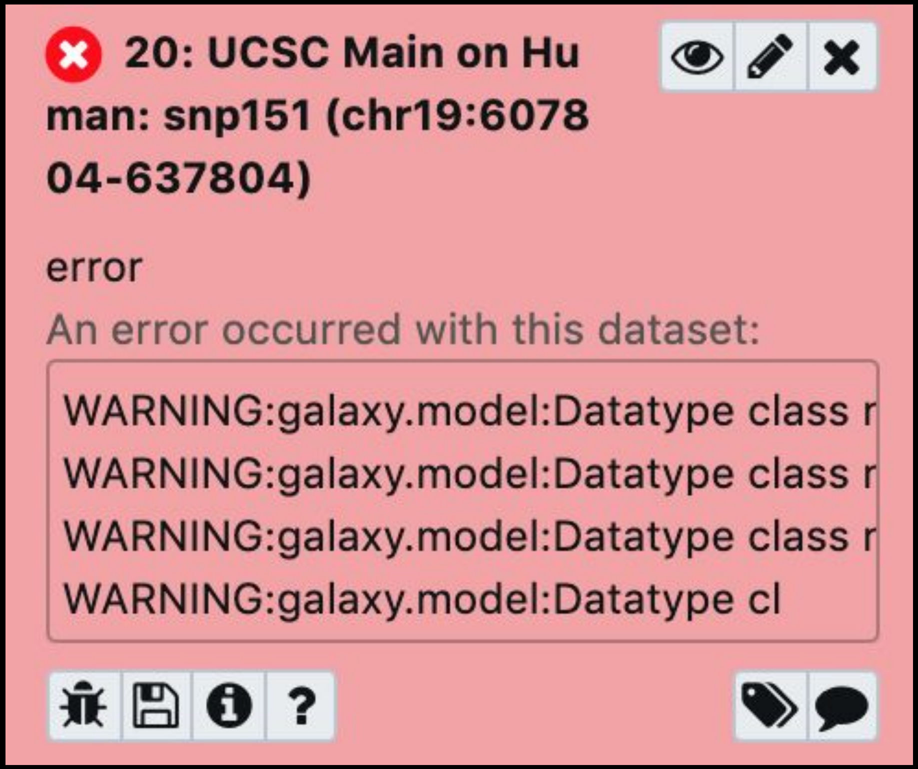
I've been trying to install package python 2.7.10 but keep getting stuck at bzlib dependency with bzlib 1.0.6 with the "make: * No targets specified and no makefile found. Stop." error. How do I fix this? Thanks!

Nicola Soranzo @nsoranzo Jun 30 23:36

Use conda!

Learning and support

- Support
  - Help.GalaxyProject.org
  - Gitter
  - **Direct reporting**

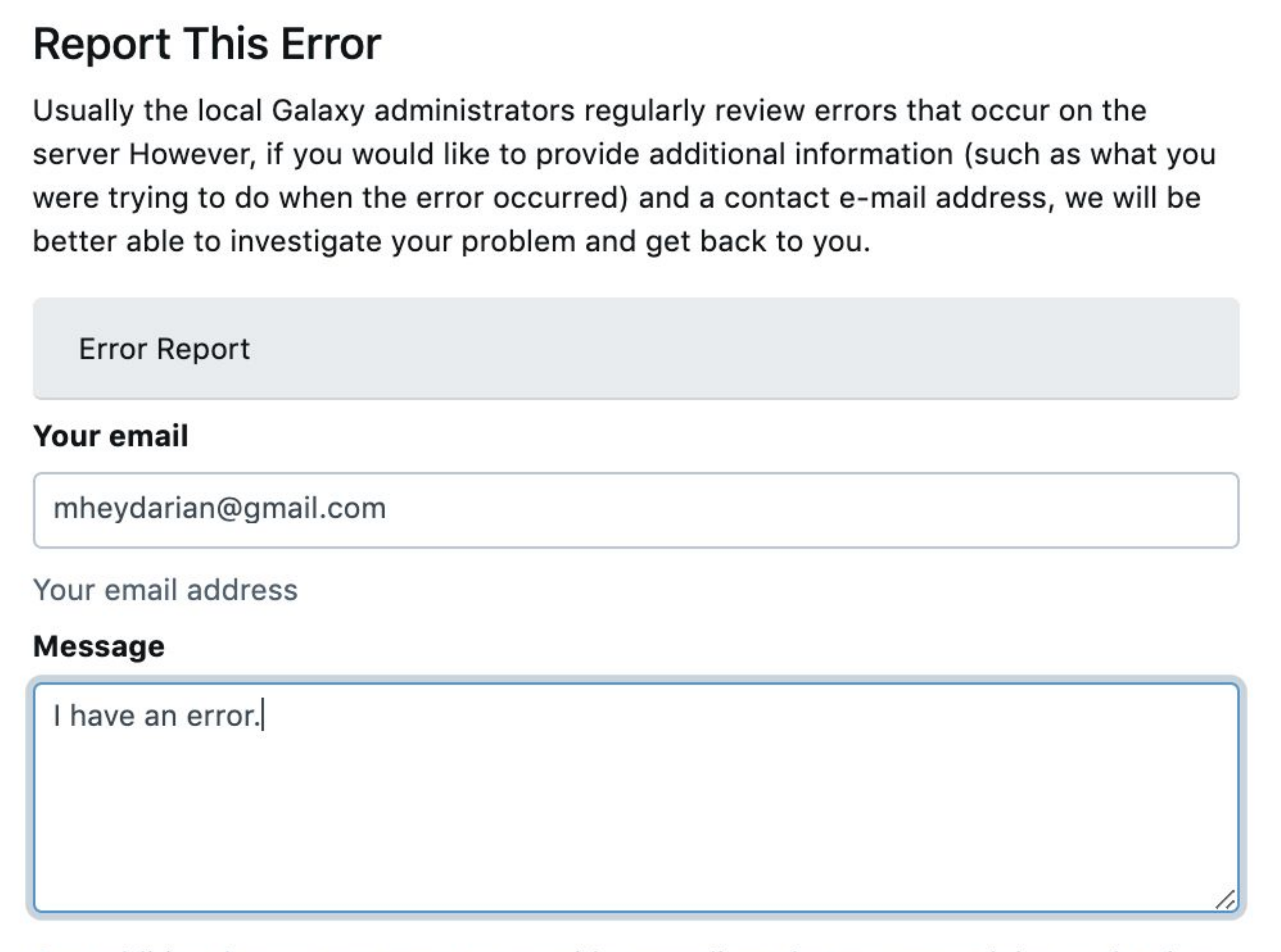


20: UCSC Main on Human: snp151 (chr19:607804-637804)

error

An error occurred with this dataset:

```
WARNING:galaxy.model:Datatype class r
WARNING:galaxy.model:Datatype class r
WARNING:galaxy.model:Datatype class r
WARNING:galaxy.model:Datatype cl
```



Report This Error

Usually the local Galaxy administrators regularly review errors that occur on the server However, if you would like to provide additional information (such as what you were trying to do when the error occurred) and a contact e-mail address, we will be better able to investigate your problem and get back to you.

Error Report

Your email

mheydarian@gmail.com

Your email address

Message

I have an error.

Any additional comments you can provide regarding what you were doing at the time

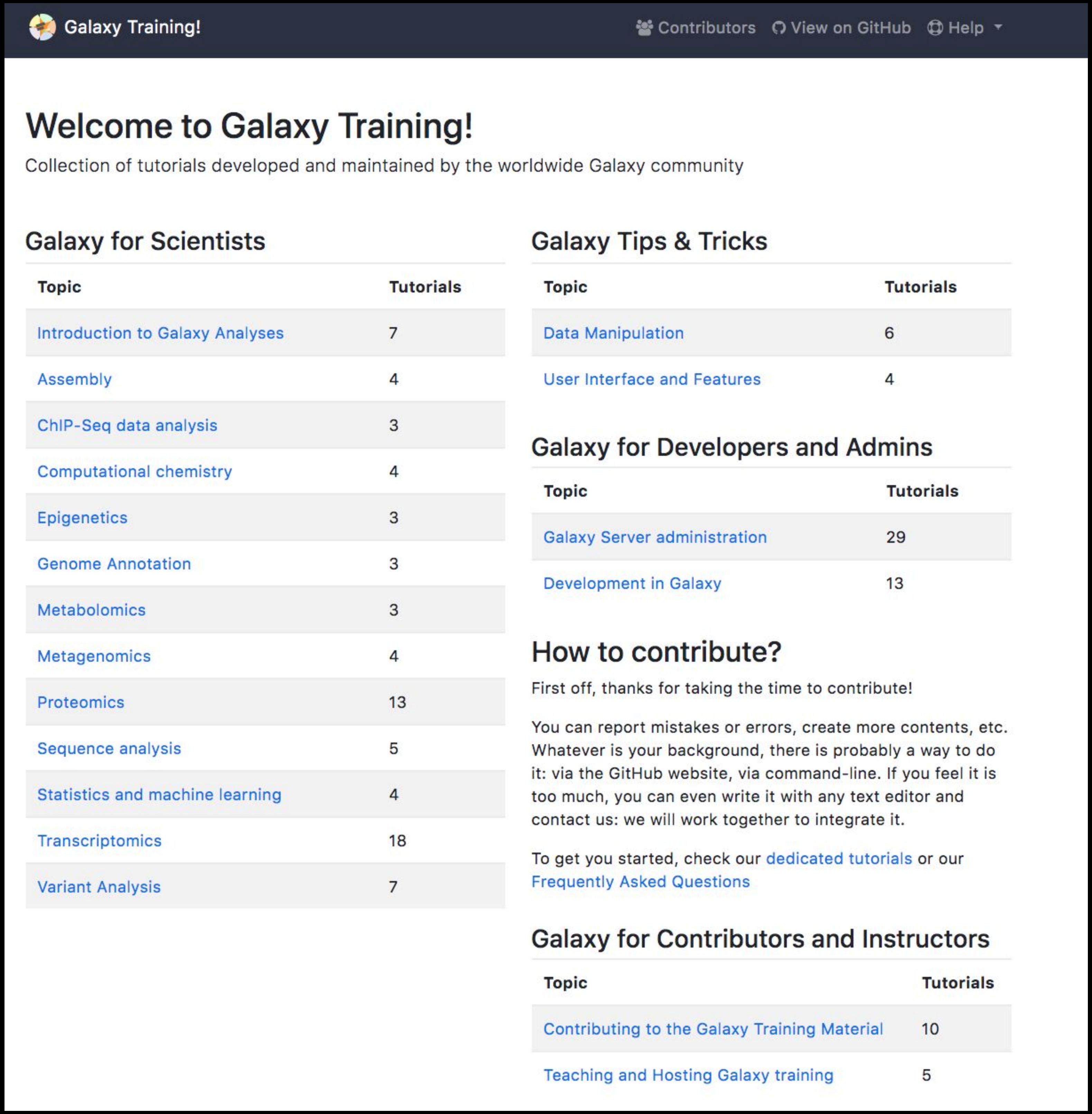
Galaxy Training Network

- Community driven
- Open source
    - https://github.com/galaxyproject/training-material
    - Include tutorials on how to customize and contribute training materials

https://training.galaxyproject.org

Galaxy Training Network

- Community based resource for educational materials using Galaxy to teach diverse domains of science.



Welcome to Galaxy Training!

Collection of tutorials developed and maintained by the worldwide Galaxy community

Galaxy for Scientists

Topic	Tutorials
Introduction to Galaxy Analyses	7
Assembly	4
ChIP-Seq data analysis	3
Computational chemistry	4
Epigenetics	3
Genome Annotation	3
Metabolomics	3
Metagenomics	4
Proteomics	13
Sequence analysis	5
Statistics and machine learning	4
Transcriptomics	18
Variant Analysis	7

Galaxy Tips & Tricks

Topic	Tutorials
Data Manipulation	6
User Interface and Features	4

Galaxy for Developers and Admins

Topic	Tutorials
Galaxy Server administration	29
Development in Galaxy	13

How to contribute?

First off, thanks for taking the time to contribute!

You can report mistakes or errors, create more contents, etc. Whatever is your background, there is probably a way to do it: via the GitHub website, via command-line. If you feel it is too much, you can even write it with any text editor and contact us: we will work together to integrate it.

To get you started, check our dedicated tutorials or our Frequently Asked Questions

Galaxy for Contributors and Instructors

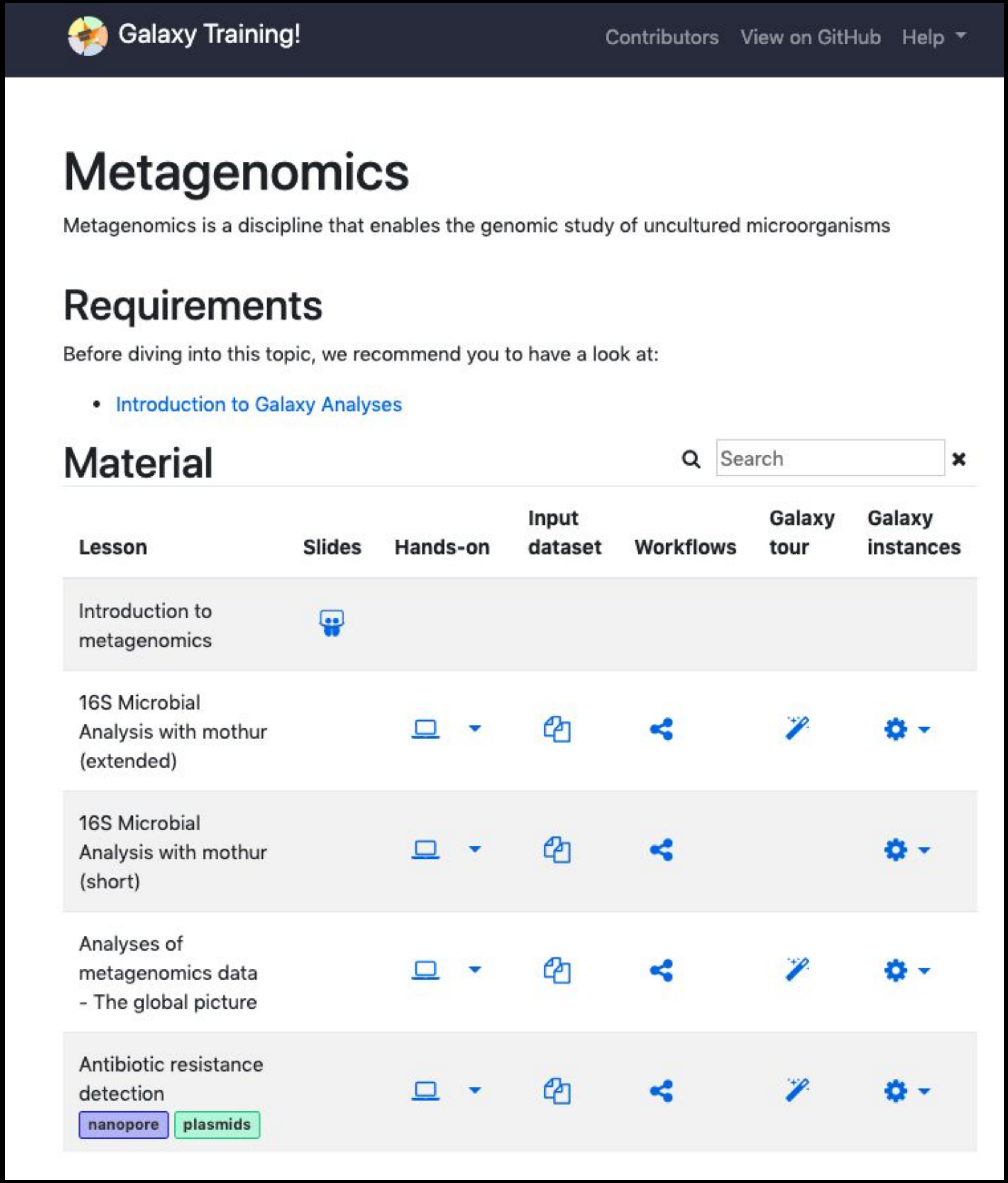
Topic	Tutorials
Contributing to the Galaxy Training Material	10
Teaching and Hosting Galaxy training	5

https://training.galaxyproject.org

Galaxy Training Network

- Tutorials are coupled with background information, sample data, and tool parameter recommendations.
- Slides, workflows, and interactive tours are included with most tutorials.
- Accessibility on Galaxy instances.
- Translation to Spanish, French, Japanese,

- “Out of the box” exercises.



Metagenomics

Metagenomics is a discipline that enables the genomic study of uncultured microorganisms

Requirements

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

- Introduction to Galaxy Analyses

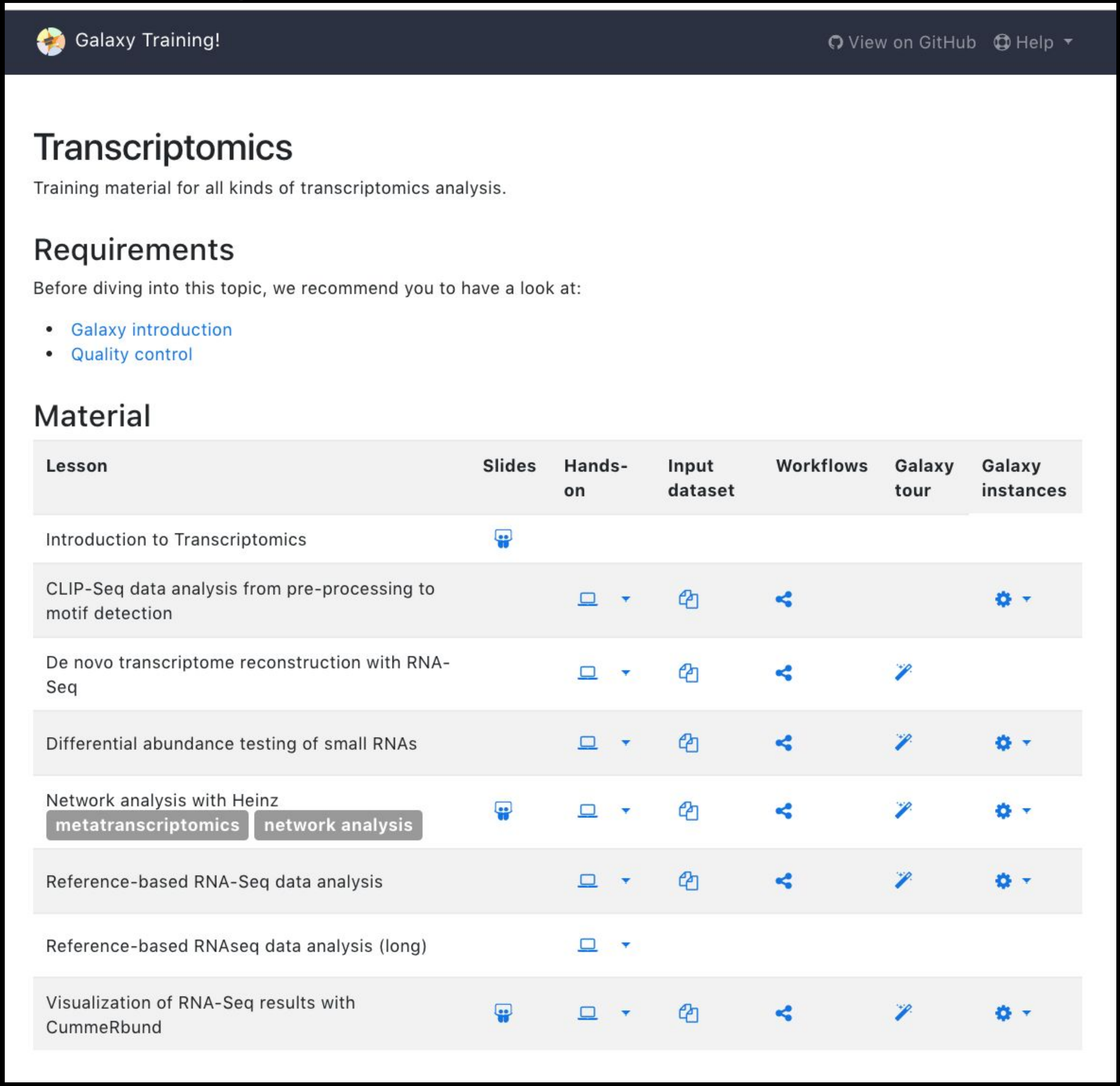
Material

Lesson	Slides	Hands-on	Input dataset	Workflows	Galaxy tour	Galaxy instances
Introduction to metagenomics	✓					
16S Microbial Analysis with mothur (extended)		✓ ▾	✓	✓	✓	✓ ▾
16S Microbial Analysis with mothur (short)		✓ ▾	✓	✓		✓ ▾
Analyses of metagenomics data - The global picture		✓ ▾	✓	✓	✓	✓ ▾
Antibiotic resistance detection nanopore plasmids		✓ ▾	✓	✓	✓	✓ ▾

https://training.galaxyproject.org

Galaxy Training Network

- Sample data
  - Downsampled to enable completion of exercises in few hours.
  - Available on Zenodo and in Data Libraries of useGalaxy.*



Transcriptomics

Training material for all kinds of transcriptomics analysis.

Requirements

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

- Galaxy introduction
- Quality control

Material

Lesson	Slides	Hands-on	Input dataset	Workflows	Galaxy tour	Galaxy instances
Introduction to Transcriptomics	🖵					
CLIP-Seq data analysis from pre-processing to motif detection		💻 ▾	⧉	<		⚙ ▾
De novo transcriptome reconstruction with RNA-Seq		💻 ▾	⧉	<	🪄	
Differential abundance testing of small RNAs		💻 ▾	⧉	<	🪄	⚙ ▾
Network analysis with Heinz metatranscriptomics network analysis	🖵	💻 ▾	⧉	<	🪄	⚙ ▾
Reference-based RNA-Seq data analysis		💻 ▾	⧉	<	🪄	⚙ ▾
Reference-based RNAseq data analysis (long)		💻 ▾				
Visualization of RNA-Seq results with CummeRbund	🖵	💻 ▾	⧉	<	🪄	⚙ ▾

https://training.galaxyproject.org

Galaxy servers

- useGalaxy.*
- Public servers
- Cloud based services
- Deploy your own

Galaxy servers

- useGalaxy.*
  - usegalaxy.org
  - usegalaxy.eu
  - usegalaxy.org.au
- Public servers
- Cloud based services
- Deploy your own

- Free to use
  - 250 GB storage/user
  - 10 concurrent jobs/user
- Significant computational resources
- Managed by system admin
- Common set of reference genomes and tools available
- Nationally/continentally funded
  - Training infrastructure as a Service (TiaaS) available on EU server

Galaxy servers
Training infrastructure as a Service (TiaaS)

- A dedicated service running on useGalaxy.EU that provides users dedicated resources to ensure educational exercises can be completed promptly.
  - Includes dashboard to monitor student usage.
- A free service.
  - https://galaxyproject.eu/tiaas
  - https://training.galaxyproject.org/training-material/topics/instructors/tutorials/setup-tiaas-for-training/tutorial.html

Galaxy servers

- useGalaxy.*
- Public servers
- Cloud based services
- Deploy your own

- Focused Galaxy instances
    - Highlight domains, tools, publications
- Variable computational resources, tool availability, and reference genomes
- Managed by system admin

Galaxy servers

- useGalaxy.*
- Public servers
- Cloud based services
- Deploy your own

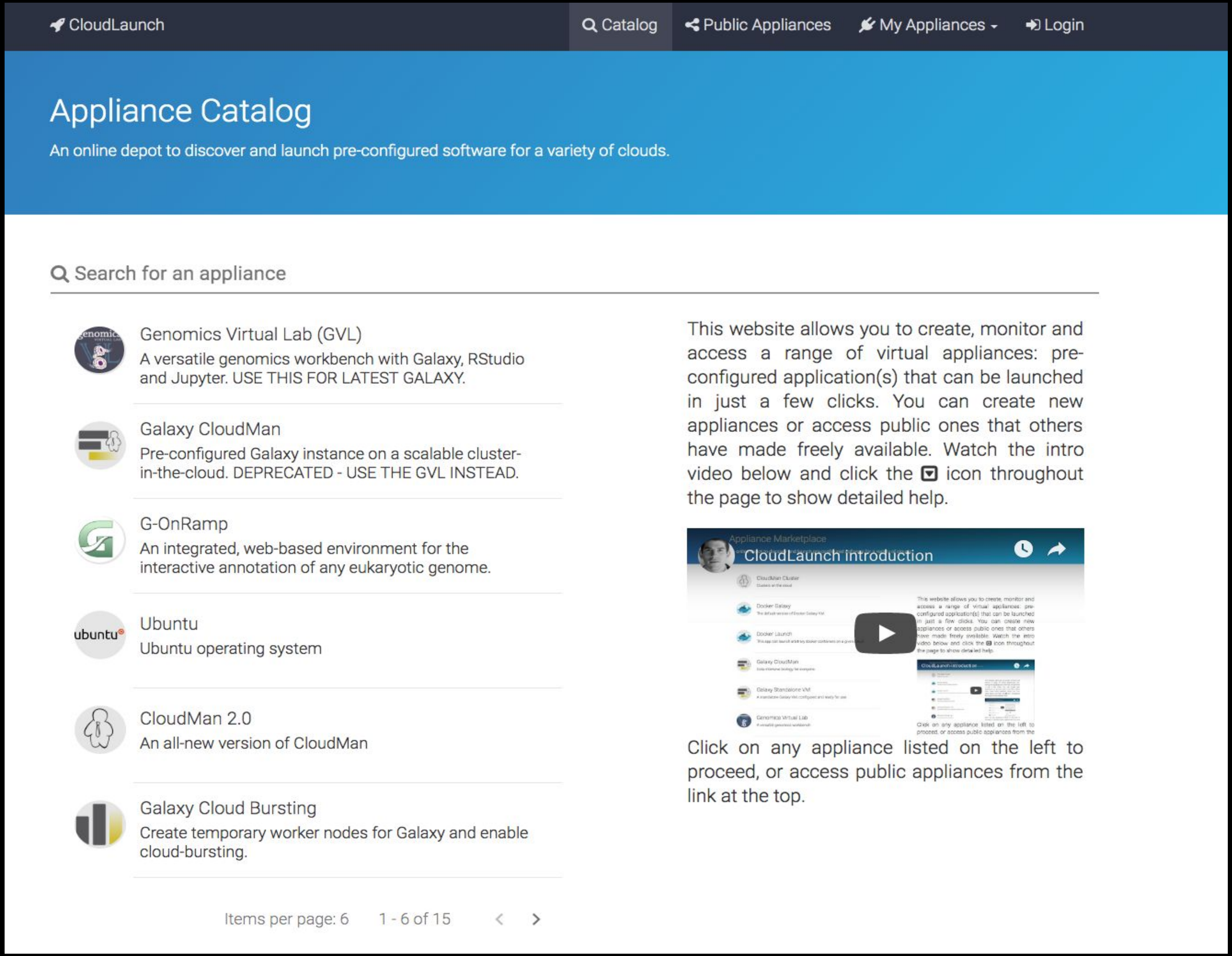
- Temporary to long term lifespan
- On demand scalability
- Commercial options: Amazon Web Services, Azure, Google Cloud Platform
- Academic options: Jetstream and Globus Genomics
- Managed by user
  - Can customize tools, reference genomes, and manage access

https://galaxyproject.org/cloud/

Galaxy on the Cloud: CloudLaunch

https://launch.usegalaxy.org/launch

- Directly launch your own Galaxy instance on AWS, Azure, Jetstream or Google Cloud Platform
- https://launch.usegalaxy.org/catalog



Appliance Catalog

An online depot to discover and launch pre-configured software for a variety of clouds.

Search for an appliance

Genomics Virtual Lab (GVL)
A versatile genomics workbench with Galaxy, RStudio and Jupyter. USE THIS FOR LATEST GALAXY.

Galaxy CloudMan
Pre-configured Galaxy instance on a scalable cluster-in-the-cloud. DEPRECATED - USE THE GVL INSTEAD.

G-OnRamp
An integrated, web-based environment for the interactive annotation of any eukaryotic genome.

Ubuntu
Ubuntu operating system

CloudMan 2.0
An all-new version of CloudMan

Galaxy Cloud Bursting
Create temporary worker nodes for Galaxy and enable cloud-bursting.

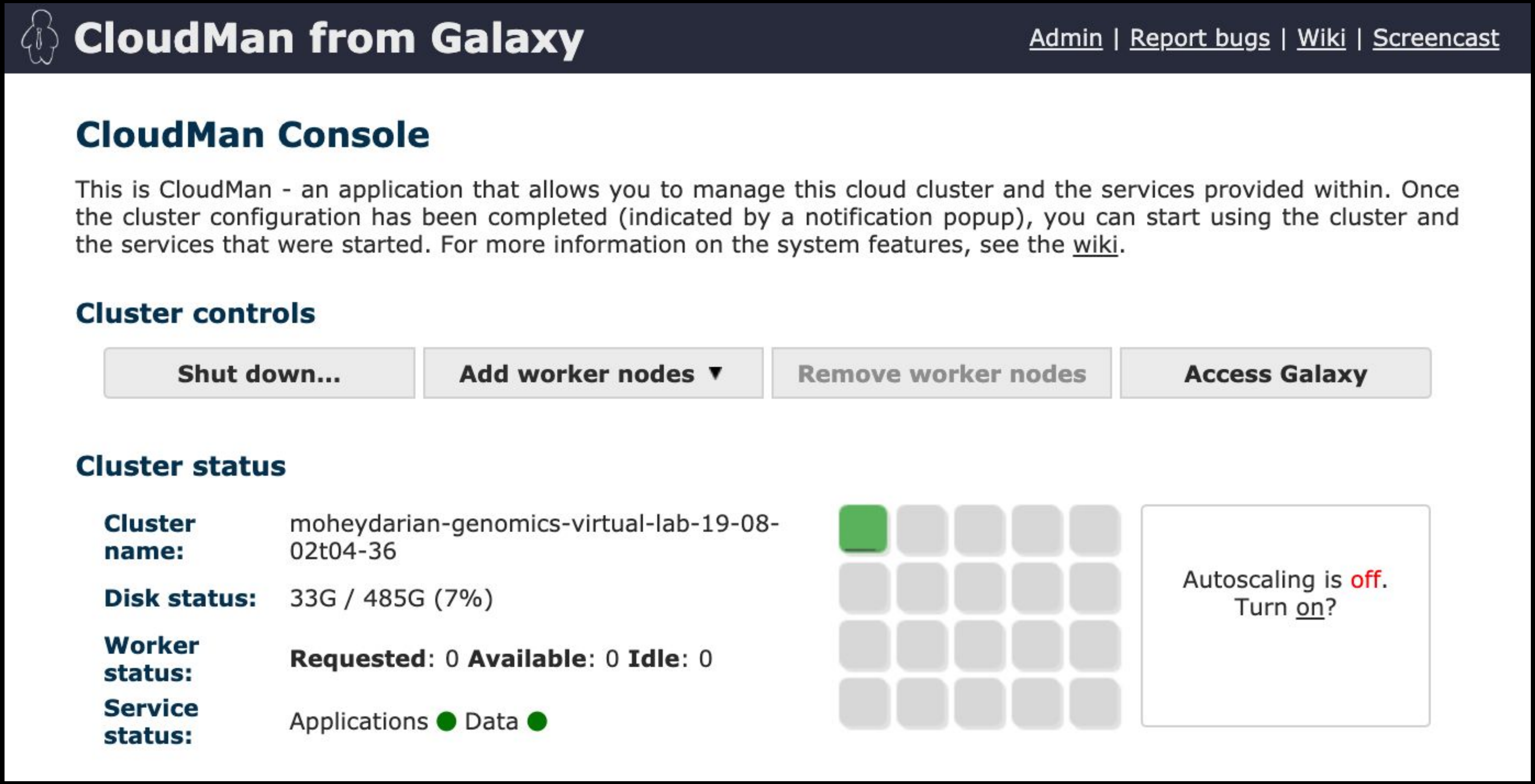
Items per page: 6 1 - 6 of 15

This website allows you to create, monitor and access a range of virtual appliances: pre-configured application(s) that can be launched in just a few clicks. You can create new appliances or access public ones that others have made freely available. Watch the intro video below and click the ⊡ icon throughout the page to show detailed help.

Click on any appliance listed on the left to proceed, or access public appliances from the link at the top.

Galaxy on the Cloud:
CloudMan for instance management

- Convenient interface for accessing and managing cloud resources



CloudMan Console

This is CloudMan - an application that allows you to manage this cloud cluster and the services provided within. Once the cluster configuration has been completed (indicated by a notification popup), you can start using the cluster and the services that were started. For more information on the system features, see the wiki.

Cluster controls

Cluster status

Cluster name: moheydarian-genomics-virtual-lab-19-08-02t04-36

Disk status: 33G / 485G (7%)

Worker status: **Requested**: 0 **Available**: 0 **Idle**: 0

Service status: Applications ● Data ●

Autoscaling is off.
Turn on?

Galaxy on the Cloud: Galaxy CloudMan

http://usegalaxy.org/cloud
http://aws.amazon.com/education

- Start with a fully configured and populated (tools and data) Galaxy instance
  - You are system admin - customize tools, ref. data, and manage access
- Allows you to scale up and down your compute assets as needed
- AWS Grants for research and education

Galaxy on the Cloud: Jetstream

- Jetstream is part of XSEDE, “a collection of advanced digital resources and services” funded by the NSF.
    - Apply for an allocation: https://portal.xsede.org/allocation-policies
- Start with a configured and populated (tools and data) Galaxy instance

Galaxy servers

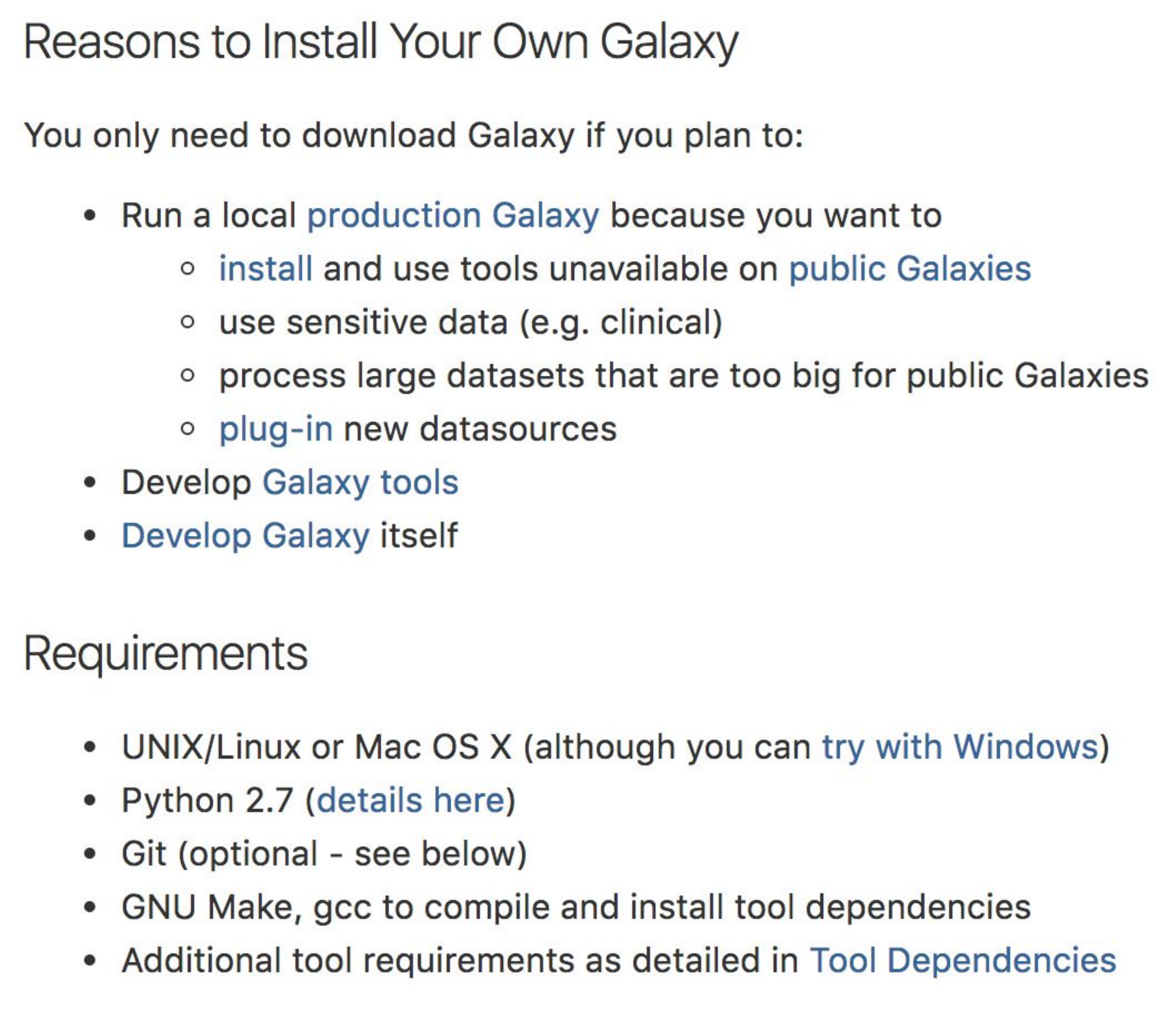
- useGalaxy.*
- Public servers
- Cloud based services
- Deploy your own
  - Base Galaxy
  - Containers
  - VMs

- Deployed on local compute infrastructure
  - Personal computer
  - Local server
- Scalability dependent on local infrastructure
- Managed by user
  - Can customize tools, reference genomes, and manage access

https://galaxyproject.org/admin/get-galaxy/

As open source software

http://getgalaxy.org



Reasons to Install Your Own Galaxy

You only need to download Galaxy if you plan to:

- Run a local production Galaxy because you want to
    - install and use tools unavailable on public Galaxies
    - use sensitive data (e.g. clinical)
    - process large datasets that are too big for public Galaxies
    - plug-in new datasources
- Develop Galaxy tools
- Develop Galaxy itself

Requirements

- UNIX/Linux or Mac OS X (although you can try with Windows)
- Python 2.7 (details here)
- Git (optional - see below)
- GNU Make, gcc to compile and install tool dependencies
- Additional tool requirements as detailed in Tool Dependencies

Benefits of administering your Galaxy

- Customize tool sets with Toolshed
  - https://toolshed.g2.bx.psu.edu/
- Customize reference genome availability
- Populate Data Libraries
- Manage users
- Eliminate the queue
- Scale with demand

Use cases for teaching with Galaxy

- Lecture setting in course
    - Demonstrate analysis pipelines in Galaxy and provide access to students

- Lab setting in course
    - Provide students access to Galaxy and GTN training materials

- Use cases above use small data for educational purposes. Can easily utilize a useGalaxy.* server to eliminate costs and server management.

Use cases for teaching with Galaxy

- Independent research or thesis project
    - Trainee would analyze actual data to answer biological questions.

- The independent research use case will likely use large (published or in-house) datasets. This can be performed on useGalaxy.*, but the trainee will hit storage quota rapidly.
- The trainee would benefit from cloud based resources.

P141: The Galaxy Training Network:
A Community Based Training Resource

Galaxy: An Open Platform for Data Analysis and Integration
Tuesday, 4pm, California Room

+ much more

https://galaxyproject.org/events/2020-pag/

Acknowledgements

You !!

Core Team

Enis Afgan
Johns Hopkins

Dannon Baker
Johns Hopkins

Dan Blankenberg
The Cleveland Clinic

Dave Bouvier
Penn State

Martin Cech
Penn State

John Chilton
Penn State

Dave Clements
Johns Hopkins

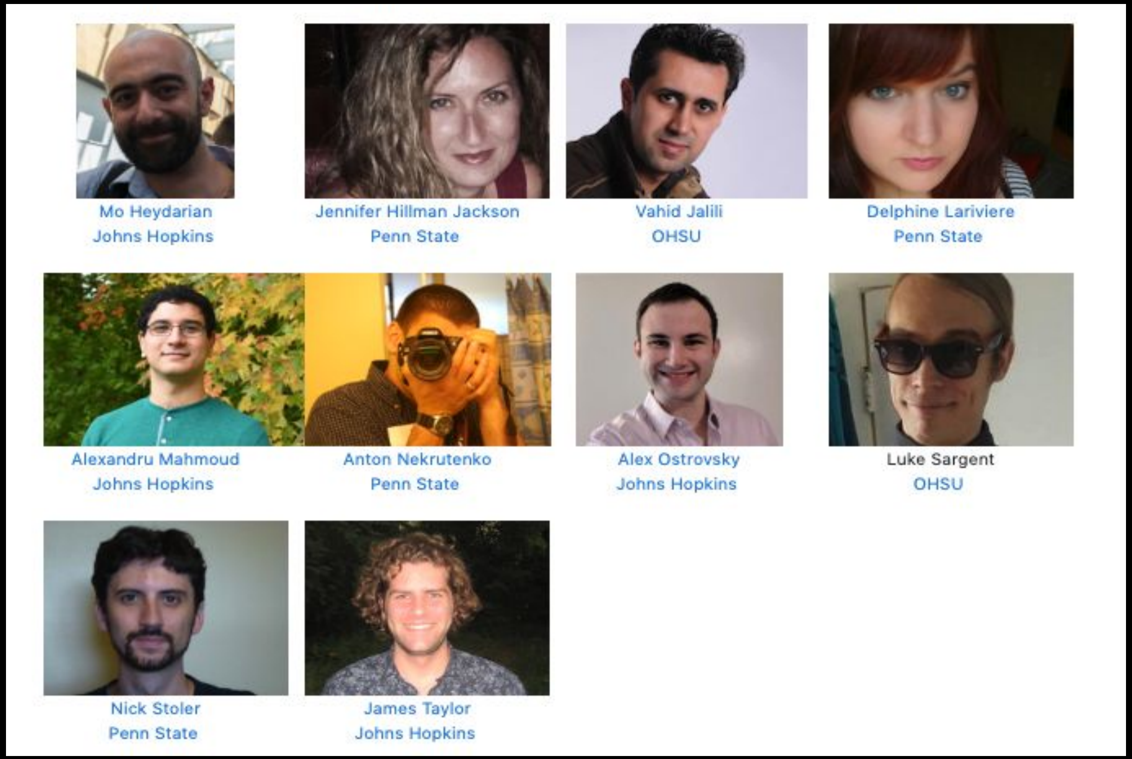
Nate Coraor
Penn State

Jeremy Goecks
OHSU

Sergey Golitsynskiy
Johns Hopkins

Qiang Gu
OHSU

Aysam Guerler
Johns Hopkins



Mo Heydarian
Johns Hopkins

Jennifer Hillman Jackson
Penn State

Vahid Jalili
OHSU

Delphine Lariviere
Penn State

Alexandru Mahmoud
Johns Hopkins

Anton Nekrutenko
Penn State

Alex Ostrovsky
Johns Hopkins

Luke Sargent
OHSU

Nick Stoler
Penn State

James Taylor
Johns Hopkins

PAG 2020 Organizers
Dr. Abdelmajid Kassem

The Galaxy Team
Freiburg Galaxy Team
European Galaxy Project
Australian Galaxy Project
National Institutes of Health
Johns Hopkins University
Penn State University
Oregon Health & Science University
Lerner Research Institute - Cleveland Clinic