

5th Global Research Platform Workshop @ eScience

Sept. 16-20, 2024, Osaka, Japan

UseGalaxy Canada

Implementing the Canadian Branch of the UseGalaxy Consortium

Carol Gauthier
Université de Sherbrooke



Université de
Sherbrooke



Digital Research
Alliance of Canada

Alliance de recherche
numérique du Canada

Summary

- Galaxy platform overview
 - Some examples and screenshots
 - Components and features
 - Training materials and Community
- The UseGalaxy.* Initiative
 - Why, Where and How
 - Leveraging distributed resources
- The UseGalaxy Canada branch

What is the Galaxy platform

- Galaxy is a free, open-source, easy to use platform for data analysis, workflow management, visualization and sharing
 - => accessible, reproducible, and collaborative sciences (*FAIR*)!
- Contains > 9,500 curated tools
- Supports > 400 types of input data, > 50 types of inbuilt visualizations
- Originally designed for life sciences but is becoming more science agnostic: Astronomy, Climate, AI, etc.
- Intuitive interface + APIs and bindings to interoperate with command line codes

What is the Galaxy platform

- Can be installed locally or accessed via hundreds of public and private instances
- *UseGalaxy.** is a federation of coordinated major installations
 - leveraging High Performance and Cloud computing
 - collectively supporting > 500,000 users
 - processing >2 million jobs monthly
- Has a large and vibrant global community of dedicated developers, maintainers and trainers (>400 tutorials)

What is the Galaxy platform

Galaxy Interface

... Some screenshots ...

List of tools

The screenshot displays the Galaxy web interface. On the left, the 'Tools' menu is visible, with a red circle highlighting the 'Testing Tool' section. The main panel shows 'Dataset Information' for a MACS2 callpeak dataset. The right panel shows a 'History' list of previous jobs.

Tools Menu (Left Panel):

- 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
- Statistics
- Graph/Display Data
- Phenotype Association
- Testing Tool**
- Imaging
- FASTA/FASTQ
- Annotation
- RNA Analysis

Dataset Information (Main Panel):

Property	Value
Number	12
Name	MACS2 callpeak on data 9 (narrow Peaks)
Created	Thursday Nov 9th 17:10:02 2023 GMT-5
Filesize	6 KB
Dbkey	sacCer3
Format	bed
File contents	contents
History Content API ID	ee240d7b45a01b08 (1477)
History API ID	d5024d270d7f028b (44)
UUID	33b9b634-df03-43a6-ac90-dfee9a7a6324
Full Path	/mnt/nfs/data/datasets/3/3/b/dataset_33b9b634-df03-43a6-ac90-dfee9a7a6324.dat

Tool Parameters (Main Panel):

Input Parameter	Value	Note for rerun
Are you pooling Treatment Files?		not used (parameter was added after this job was run)
Do you have a Control File?		not used (parameter was added after this job was run)
Format of Input Files		not used (parameter was added after this job was run)
effective_genome_size_options		The previously used value is no longer valid.
Build Model	nomodel	
Set extension size	100	
Set shift size	0	
Peak detection based on	qvalue	
Minimum FDR (q-value) cutoff for peak detection	1e-10	
Additional Outputs	Peaks as tabular file (compatible with MultiQC) Peak summits Scores in bedGraph files (--bdg)	
advanced_options		
When set, scale the small sample		not used (parameter was added after this job was run)

History (Right Panel):

- 16: Wig/BedGraph-to-bigWig on data 14
- 15: MACS2 callpeak on data 9 (Bedgraph Control)
- 14: MACS2 callpeak on data 9 (Bedgraph Treatment)
- 13: MACS2 callpeak on data 9 (summits in BED)
- 12: MACS2 callpeak on data 9 (narrow Peaks)
- 11: MACS2 callpeak on data 9 (Peaks in tabular format)
- 10: Flagstat on data 9
- 9: Filter SAM or BAM, output SAM or BAM on data 7: bam
- 8: Flagstat on data 7

Tool parameters and/or Dataset information

The screenshot displays the Galaxy web interface. On the left is a sidebar with navigation links. The main area is divided into two panels: 'Dataset Information' and 'Tool Parameters'. A red circle highlights these two panels. The 'Dataset Information' panel shows details for a dataset named 'MACS2 callpeak on data 9 (narrow Peaks)'. The 'Tool Parameters' panel shows input parameters for the MACS2 tool, with some parameters marked as 'not used' or 'no longer valid'. On the right, a 'History' panel shows a list of recent jobs.

Dataset Information

Property	Value
Number	12
Name	MACS2 callpeak on data 9 (narrow Peaks)
Created	Thursday Nov 9th 17:10:02 2023 GMT-5
Filesize	6 KB
Dbkey	sacCer3
Format	bed
File contents	contents
History Content API ID	ee240d7b45a01b08 (1477)
History API ID	d5024d270d7f028b (44)
UUID	33b9b634-df03-43a6-ac90-dfee9a7a6324
Full Path	/mnt/nfs/data/datasets/3/3/b/dataset_33b9b634-df03-43a6-ac90-dfee9a7a6324.dat

Tool Parameters

Input Parameter	Value	Note for rerun
Are you pooling Treatment Files?		not used (parameter was added after this job was run)
Do you have a Control File?		not used (parameter was added after this job was run)
Format of Input Files		not used (parameter was added after this job was run)
effective_genome_size_options		The previously used value is no longer valid.
Build Model	nomodel	
Set extension size	100	
Set shift size	0	
Peak detection based on	qvalue	
Minimum FDR (q-value) cutoff for peak detection	1e-10	
Additional Outputs	Peaks as tabular file (compatible with MultiQC) Peak summits Scores in bedGraph files (--bdg)	
advanced_options		
When set, scale the small sample		not used (parameter was added after this job was run)

History

search datasets

BFT708_cours2

277 MB 16

- 16: Wig/BedGraph-to-bigWig on data 14
- 15: MACS2 callpeak on data 9 (Bedgraph Control)
- 14: MACS2 callpeak on data 9 (Bedgraph Treatment)
- 13: MACS2 callpeak on data 9 (summits in BED)
- 12: MACS2 callpeak on data 9 (narrow Peaks)
- 11: MACS2 callpeak on data 9 (Peaks in tabular format)
- 10: Flagstat on data 9
- 9: Filter SAM or BAM, output SAM or BAM on data 7: bam
- 8: Flagstat on data 7

History to keep full records of analysis operations

The screenshot displays the Galaxy web interface. On the left is a sidebar with a 'Tools' section containing a search bar and an 'Upload Data' button, followed by a list of tool categories: 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, Statistics, Graph/Display Data, Phenotype Association, Testing Tool, Imaging, FASTA/FASTQ, Annotation, and RNA Analysis. The main area is divided into two panels. The left panel, 'Dataset Information', shows details for a dataset named 'MACS2 callpeak on data 9 (narrow Peaks)', including its number (12), creation date, file size (6 KB), and full path. The right panel, 'Tool Parameters', lists input parameters and their values for the same dataset. The rightmost panel, 'History', is circled in red and shows a list of 16 analysis jobs. Job 12, 'MACS2 callpeak on data 9 (narrow Peaks)', is highlighted in blue, indicating it is the current job. Other jobs include 'Wig/BedGraph-to-bigWig on data 14', 'MACS2 callpeak on data 9 (Bedgraph Control)', 'MACS2 callpeak on data 9 (Bedgraph Treatment)', 'MACS2 callpeak on data 9 (summits in BED)', 'MACS2 callpeak on data 9 (Peaks in tabular format)', 'Flagstat on data 9', 'Filter SAM or BAM, output SAM or BAM on data 7: bam', and 'Flagstat on data 7'.

Dataset Information

Number	12
Name	MACS2 callpeak on data 9 (narrow Peaks)
Created	Thursday Nov 9th 17:10:02 2023 GMT-5
Filesize	6 KB
Dbkey	sacCer3
Format	bed
File contents	contents
History Content API ID	ee240d7b45a01b08 (1477)
History API ID	d5024d270d7f028b (44)
UUID	33b9b634-df03-43a6-ac90-dfee9a7a6324
Full Path	/mnt/nfs/data/datasets/3/3/b/dataset_33b9b634-df03-43a6-ac90-dfee9a7a6324.dat

Tool Parameters

Input Parameter	Value	Note for rerun
Are you pooling Treatment Files?		not used (parameter was added after this job was run)
Do you have a Control File?		not used (parameter was added after this job was run)
Format of Input Files		not used (parameter was added after this job was run)
effective_genome_size_options		The previously used value is no longer valid.
Build Model	nomodel	
Set extension size	100	
Set shift size	0	
Peak detection based on	qvalue	
Minimum FDR (q-value) cutoff for peak detection	1e-10	
Additional Outputs	Peaks as tabular file (compatible with MultiQC) Peak summits Scores in bedGraph files (--bdg)	
advanced_options		
When set, scale the small sample		not used (parameter was added after this job was run)

History

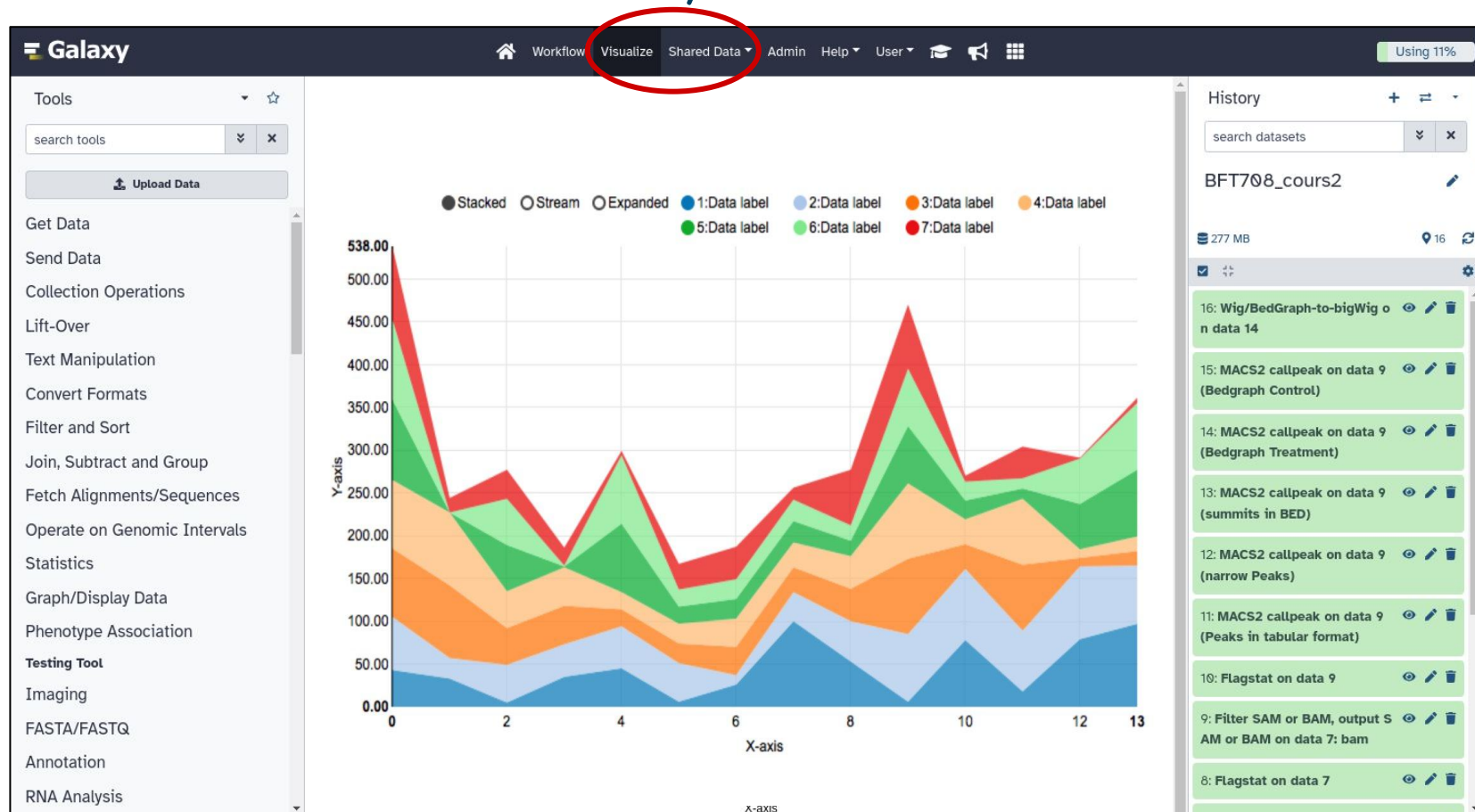
- 16: Wig/BedGraph-to-bigWig on data 14
- 15: MACS2 callpeak on data 9 (Bedgraph Control)
- 14: MACS2 callpeak on data 9 (Bedgraph Treatment)
- 13: MACS2 callpeak on data 9 (summits in BED)
- 12: MACS2 callpeak on data 9 (narrow Peaks)
- 11: MACS2 callpeak on data 9 (Peaks in tabular format)
- 10: Flagstat on data 9
- 9: Filter SAM or BAM, output SAM or BAM on data 7: bam
- 8: Flagstat on data 7

Exploring data

The screenshot displays the Galaxy web interface. On the left is a sidebar with tool categories: 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, Statistics, Graph/Display Data, Phenotype Association, Testing Tool, Imaging, FASTA/FASTQ, Annotation, and RNA Analysis. The main area shows a table with 10 columns and 22 rows of genomic data. The right sidebar contains a 'History' panel listing 16 datasets, with the 12th dataset, 'MACS2 callpeak on data 9 (narrow Peaks)', highlighted by a red circle.

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10
chrI	71227	71328	MACS2_peak_1	227	.	4.58015	25.99300	22.71520	56
chrI	139600	139704	MACS2_peak_2	151	.	4.19968	18.11684	15.11145	73
chrI	141785	141942	MACS2_peak_3	526	.	7.99677	56.30179	52.67088	83
chrI	166239	166346	MACS2_peak_4	213	.	6.18818	24.56419	21.32440	67
chrII	197457	197561	MACS2_peak_5	259	.	6.42303	29.32169	25.97875	46
chrII	326796	326919	MACS2_peak_6	235	.	6.45963	26.80078	23.50461	59
chrII	431111	431357	MACS2_peak_7	447	.	7.06150	48.35646	44.78230	142
chrII	798763	799053	MACS2_peak_8	753	.	9.87854	79.17228	75.37743	158
chrIII	138596	138922	MACS2_peak_9	180	.	4.39560	21.18610	18.06924	91
chrIII	163595	163767	MACS2_peak_10	245	.	6.05428	27.88711	24.57026	86
chrIII	197198	197412	MACS2_peak_11	1223	.	16.44144	126.57542	122.30856	97
chrIII	227883	228013	MACS2_peak_12	275	.	6.16570	30.88861	27.52250	47
chrIII	298994	299335	MACS2_peak_13	1190	.	13.70338	123.26751	119.05589	135
chrIV	273879	274016	MACS2_peak_14	340	.	7.85276	37.48251	34.02022	46
chrIV	481780	481947	MACS2_peak_15	668	.	10.24390	70.56670	66.83142	80
chrIV	539198	539415	MACS2_peak_16	944	.	14.47527	98.43713	94.48912	126
chrIV	556062	556174	MACS2_peak_17	123	.	4.13534	15.20779	12.34871	77
chrIV	1022483	1022687	MACS2_peak_18	1066	.	13.83399	110.66152	106.63006	94
chrIV	1175822	1175922	MACS2_peak_19	201	.	5.72738	23.36007	20.16481	72
chrIV	1345348	1345486	MACS2_peak_20	388	.	8.17031	42.31963	38.80029	65
chrIV	1420689	1420885	MACS2_peak_21	606	.	9.53289	64.32202	60.62062	102
chrM	16356	16456	MACS2_peak_22	180	.	5.68302	21.21186	18.08946	50

Visualize and share data analysis



Build, manage and share Workflows

Galaxy Workflow Visualize Shared Data Admin Help User

Using 11%

Tools

search tools

Upload Data

Inputs

Data Managers

Get Data

Send Data

Collection Operations

Expression Tools

Lift-Over

Text Manipulation

Convert Formats

Filter and Sort

Join, Subtract and Group

Fetch Alignments/Sequences

Operate on Genomic Intervals

Statistics

Graph/Display Data

Phenotype Association

Testing Tool

Imaging

Compute daily and monthly mean from meteorological station measurements (imported from uploaded file)

1 Temperature (timeseries (input))

2 Convert CSV to tabular (tabular)

3 Design data columns to YYYYMMDD (File to process, tabular (input))

4 Filter (Input tabular dataset, %out_file (input))

5 Sequence only (Filter, %out_file (input))

6 Positive only (Filter, %out_file (input))

7 Daily mean (Input tabular dataset, %out_file (input))

8 Monthly mean (Input tabular dataset, %out_file (input))

9 Daily climate (timeseries to plot, output_daily_temperature (input))

10 Monthly climate (timeseries to plot, output_monthly_temperature (input))

11 Daily climate (timeseries to plot, output_daily_temperature (input))

12 Monthly climate (timeseries to plot, output_monthly_temperature (input))

climate stripes
from timeseries
(Galaxy Version 1.0.2)

Label

Monthly climate stripes

Add a step label.

Step Annotation

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

Conditionally skip step?

☐ No

Set to true and connect a boolean parameter that determines whether the step should run. The step runs if the parameter value is true and will be skipped if the parameter value is false

Tool Parameters

timeseries to plot *

Data input 'filename' (tabular)

☐ column name to use for plotting - optional

Galaxy

... Some more details ...

Tools, Toolshed and Apptainer

- Tools are a central part of Galaxy
 - 9000+ tools already available in the toolshed
(<https://toolshed.g2.bx.psu.edu/>)
 - Developed by the community using a variety of languages and libraries
 - There are utilities to assist in developing and deploying tools
 - Planemo: <https://planemo.readthedocs.io/en/latest/readme.html>
- Tools are now packaged in containers (Apptainer/Docker)
 - Managing tools installations and dependencies is time consuming
 - The vast majority of tools have already been packaged in containers
 - Distributed through CVMFS
 - Registry of tools prepackaged: <https://biocontainers.pro/>

Galaxy API and BioBlend

- **Web interface is good for:**
 - Exploratory analysis and construction of Workflows
 - Visualization
 - Running workflows with new datasets
- **Galaxy has an extensive API:**
 - Connecting Galaxy directly to your data source
 - Running Workflows against a large number of datasets
 - <https://galaxyproject.org/develop/api/>
- **Language Bindings and CLI:**
 - BioBlend: Python bindings to galaxy API (<https://bioblend.readthedocs.io>)
 - Parsec: CLI tool to use Galaxy from command line
 - Other bindings for Java and PHP languages

Galaxy ... for science communities of all disciplines!

- The Toolshed already has hundreds of tools NOT specifically for genomics or life sciences
 - Statistics (514)
 - Text Manipulation (188)
 - Machine Learning (69)
 - Astronomy (11)
 - GIS (Geographic Information Systems) (4)
 - Visualization (253)
 - Computational chemistry (187)
 - Convert Format (152)
 - Data export (20)
 - Constructive Solid Geometry (11)
 - Climate analysis (12)
 - Muon Spectroscopy (10)
- Several initiatives for disciplines other than genomics related
 - Galaxy imaging: <https://imaging.usegalaxy.eu/>
 - Galaxy Machine learning: <https://ml.usegalaxy.eu/>
 - Galaxy for ecology: <https://ecology.usegalaxy.eu/>

Galaxy

... Training materials ...

Galaxy Training Network (GTN)

<https://training.galaxyproject.org/>

Welcome to Galaxy Training!

Collection of tutorials developed and maintained by the worldwide Galaxy community

Galaxy for Scientists

We have separated the tutorials into several categories based on field and technology. We are exploring other ways to organise the tutorials going forward!

Start Here

Topic	Tutorials
Introduction to Galaxy Analyses	13
Using Galaxy and Managing your Data	23

Not sure where to start?

Try the NGS Basics Learning Path!

[Start Learning](#)

Scientific Fields

Topic	Tutorials
Climate	11
Computational chemistry	9
SARS-CoV-2	9
Foundations of Data Science	49

Quickstart

Learning Pathways



Galaxy for
SysAdmins



Galaxy for
Developers



Galaxy for Teachers



Upcoming Events

Check out upcoming events around the Galaxy!

September 2 - 3, 2024

EuroBioC2024 Carpentries workshops

September 10 - 13, 2024

A practical introduction to bioinformatics and RNA-seq using Galaxy

November 12 - 14, 2024

BioNT Community Event & CarpentryConnect - Heidelberg 2024

GTN Statistics

33

Topics

408

Tutorials

18

Learning Paths

441

FAQs

373

Contributors

9.1

Years

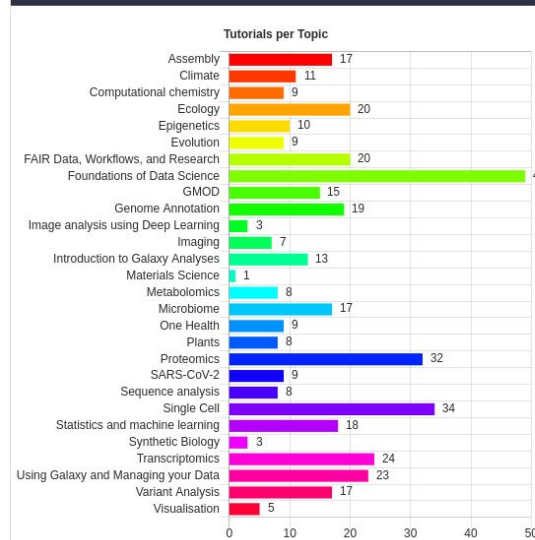
96

News Posts

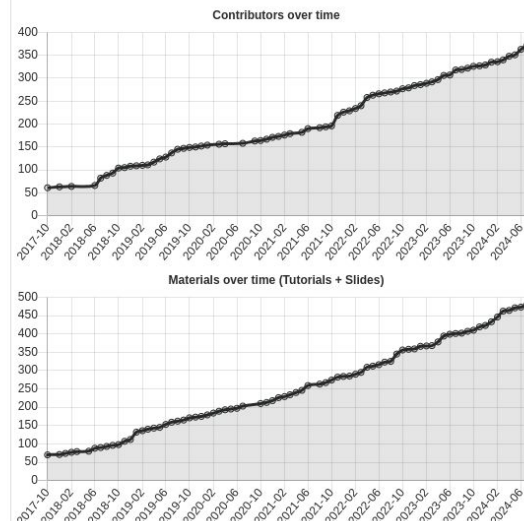
159

Videos (108.4h)

28 Scientific Topics



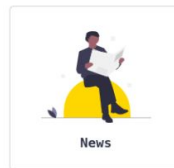
Growing Community



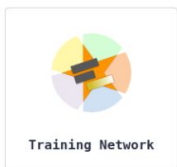
For more information about the Galaxy Project



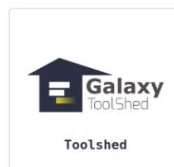
<https://galaxyproject.org/>



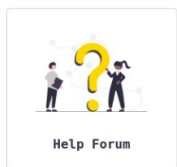
<https://galaxyproject.org/news/>



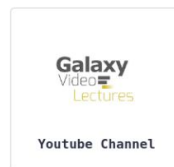
<https://training.galaxyproject.org/>



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



<https://help.galaxyproject.org/>



<https://www.youtube.com/@GalaxyProject>

UseGalaxy

vs

Galaxy

What are UseGalaxy servers

The [UseGalaxy](#) servers are [publicly accessible Galaxy servers](#) that [will support](#) a common and synchronized set of tools and reference genomes. These servers have significant computational resources behind them and are capable of powering large user communities. Some may have a regional focus, but they are all accessible to anyone.



UseGalaxy.org

[UseGalaxy.org](#) is the Galaxy Project's publicly accessible server. It was the original Galaxy server and has over 100,000 registered users, and is supported by significant computational resources at the Texas Advanced Computing Center (TACC). [Playbooks](#)

UseGalaxy.eu

[UseGalaxy.eu](#) is a publicly accessible server that aims to support the analysis and sharing needs of the European research community. It was launched in March 2018 at European Galaxy User Meeting in Freiburg. It is hosted at the University of Freiburg. [Playbooks](#)

UseGalaxy.org.au

[UseGalaxy.org.au](#) aims to support the Australian research community. It is built on the Genomics Virtual Laboratory infrastructure, and was also launched in March 2018. This server was formerly the University of Queensland Galaxy Server. [Playbooks](#)

UseGalaxy.fr

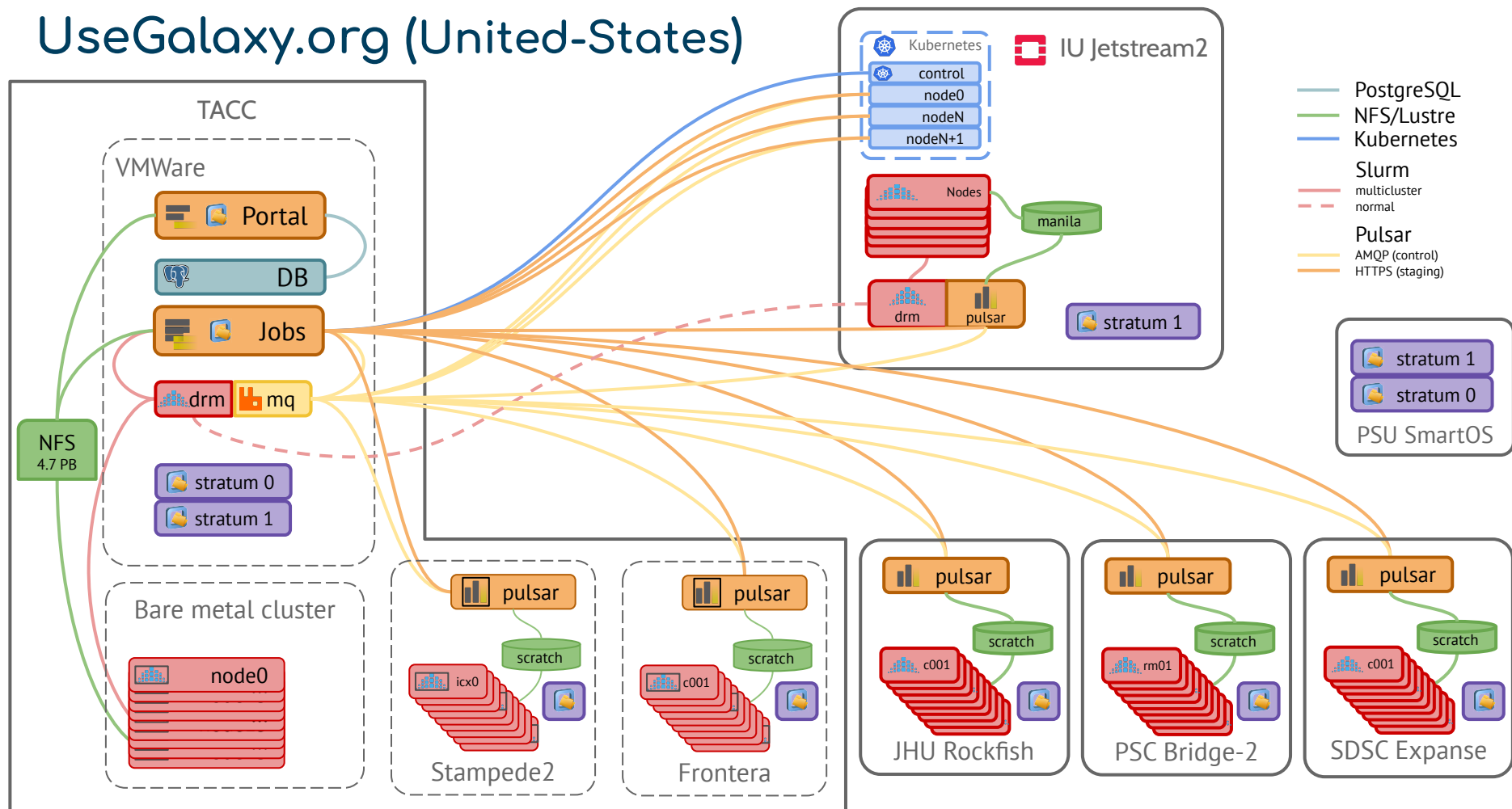
[UseGalaxy.fr](#) is a publicly accessible server aiming to support the French research community. [Playbooks](#)

UseGalaxy.ca

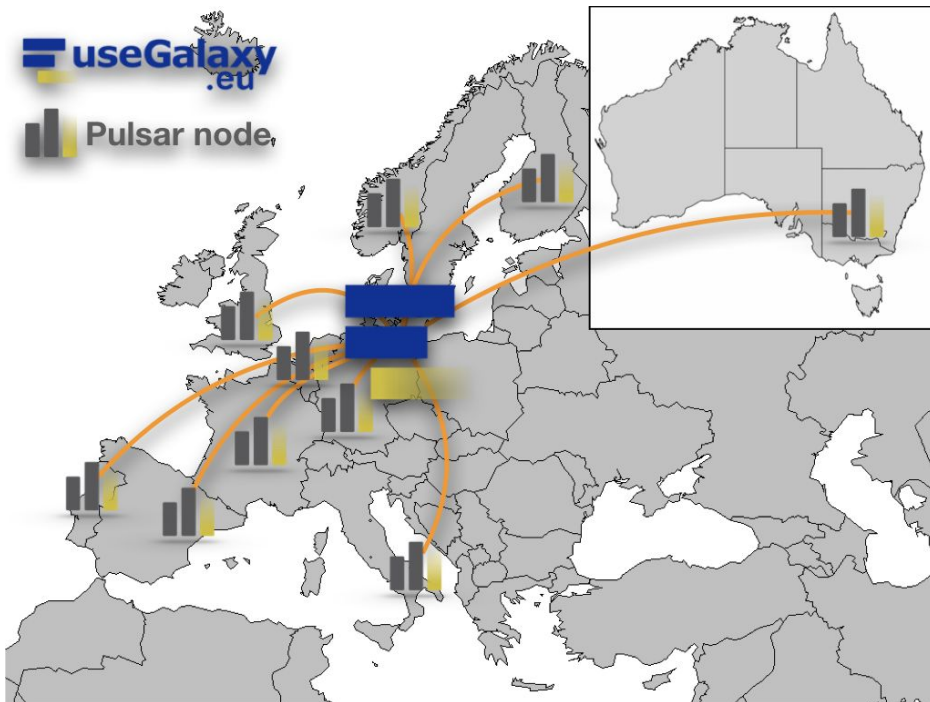
2024-25

[UseGalaxy.ca](#) is a publicly accessible server aiming to support the Canadian research community. It is supported by large Cloud resources at Calcul Québec and the Digital Research Alliance of Canada

UseGalaxy.org (United-States)

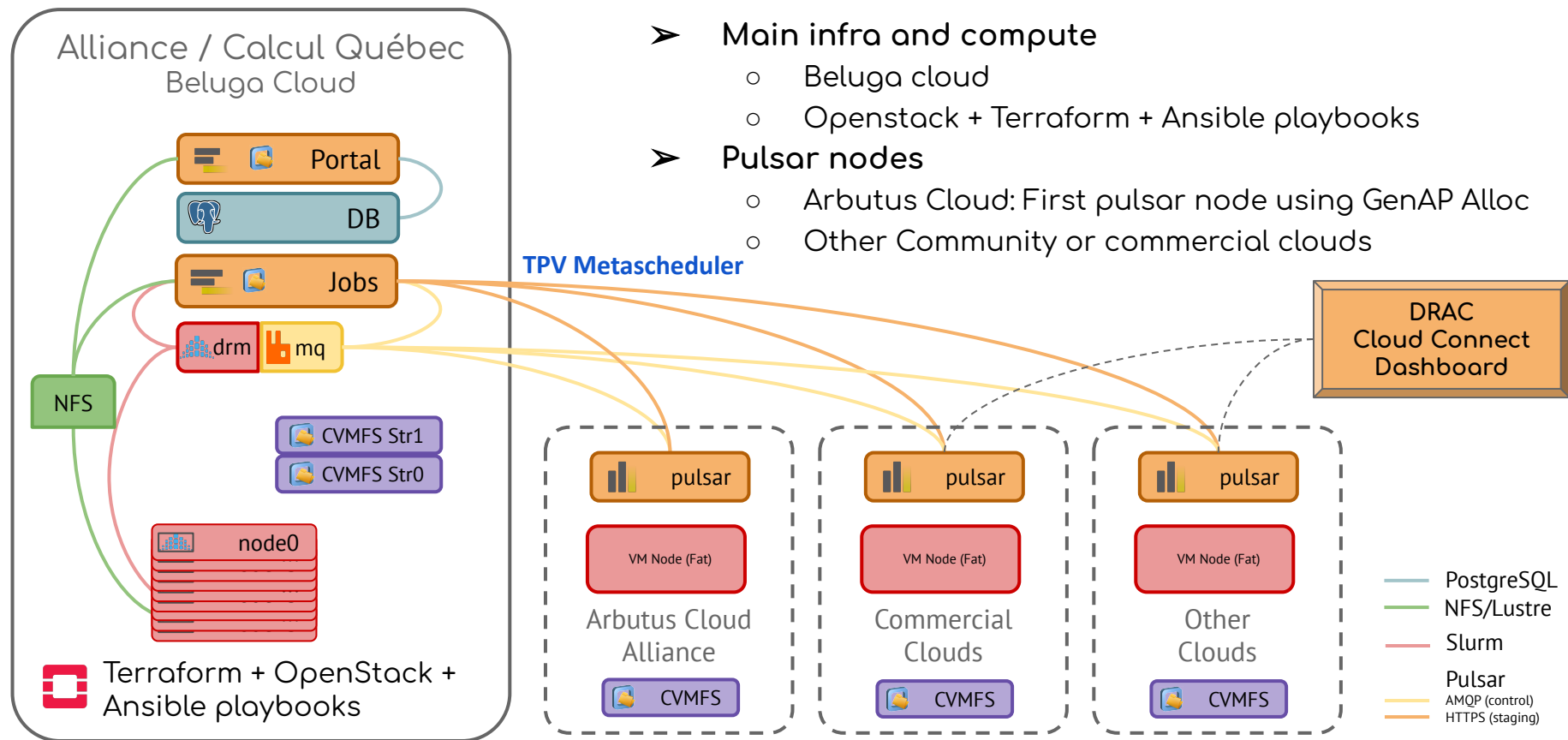


Distributed resources in *UseGalaxy*.*



- Pulsar remote computing nodes
 - Pulls jobs from the UseGalaxy server
 - Run jobs on remote infrastructure (Nodes, clusters, GPUs, etc..)
 - Push back results to the UseGalaxy server
- Galaxy TPV Meta-scheduler
 - Route jobs according to sets of attributes:
 - Tools requirements (Memory, CPUs, GPUs, etc..)
 - Users, groups, etc.

UseGalaxy Canada: The infrastructure



- **Main infra and compute**
 - Beluga cloud
 - Openstack + Terraform + Ansible playbooks
- **Pulsar nodes**
 - Arbutus Cloud: First pulsar node using GenAP Alloc
 - Other Community or commercial clouds

UseGalaxy Canada: Specific features & resources

- Mandatory SSO authentication through CILogon
- All tools jobs are run in containers
- ~50 compute nodes on the Beluga cloud
 - ~500 vCPUs
 - ~3 TB RAM
 - 100 TB total storage
- Each Canadian user receives 500 GB of storage
- Each International academic user receives 250 GB of storage
- All other users (Google auth, etc.) receives 5 GB of storage
- Storage backup on Calcul Québec tape archive
- French version of the public facing pages (to come ...)

UseGalaxy.ca ... in a



- Leverage a mature and widely used platform of **15+ years** in the making
 - Tens of thousands of users globally... with **10000+ active users** in the 3+ UseGalaxy.*
 - Hundreds of public and private instances of Galaxy in the World
 - Galaxy is **cited 11000+ times** in scientific publications
- Join a vibrant community of developers, administrators and trainers
 - **330+ contributors** to the **Galaxy code base**
 - **370+ contributors** to the **training materials**
 - **200+ contributors** to the **Toolshed, Community Hub, Playbooks**, etc.
- Offer a **General Purpose Science Gateway** to the Alliance ARC resources
 - Run workflows on Clouds or HPC clusters at the Alliance
 - Take advantage of 9000+ existing tools packaged in containers
 - ... And more

UseGalaxy Canada Team



Pierre-Étienne Jacques
Lead Researcher
Université de Sherbrooke
Quebec, Canada



Carol Gauthier
Project manager & Developer
Université de Sherbrooke
Quebec, Canada



Jonathan Laperle
Lead developer
Université de Sherbrooke
Quebec, Canada



Ata Roudgar
Analyst & System administrator
Simon Fraser University
BC, Canada



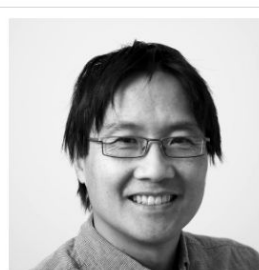
Tannistha Nandi
Bioinformatician & Data
scientist
University of Calgary
Alberta, Canada



Alain Veilleux
Director
Centre de calcul scientifique
Université de Sherbrooke
Quebec, Canada



Michel Barrette
System administrator
Université de Sherbrooke
Quebec, Canada



Minh-Nghia Nguyen
System administrator (Storage &
Backups)
Université de Sherbrooke
Quebec, Canada