Galaxy: An Open Platform for Data Analysis and Integration

PAG XXVIII, January 2020

Dave Clements,
Mo Heydarian
Johns Hopkins
University

Mathias Lorieux
CIAT

Kenneth McNally
IRRI

Star Yanxin Gao
Cornell University

Agenda

4:00	**Introduction to Galaxy and the Galaxy Ecosystem**	Dave Clements
4:20	**Galaxy for Excellence in Breeding**	Star Yanxin Gao
4:40	**Demo: GWAS with Excellence in Breeding Tools**	Mathias Lorieux, Kenneth McNally, Dave Clements, Mo Heydarian, Star Yanxin Gao
5:25	**Demo: Genomic Selection with Excellence in Breeding Tools**	Star Yanxin Gao

Introduction to Galaxy and the Galaxy Ecosystem

Plant and Animal Genome XXVIII
(PAG 2020)
San Diego, California, United States
January 14, 2020

Dave Clements, Mo Heydarian
Johns Hopkins University

#usegalaxy
@galaxyproject

Slides: bit.ly/gxy-pag-2020

Galaxy Project Outreach Team

Mo Heydarian

Johns Hopkins University
mo@galaxyproject.org
@MoHeydarian

Biology!

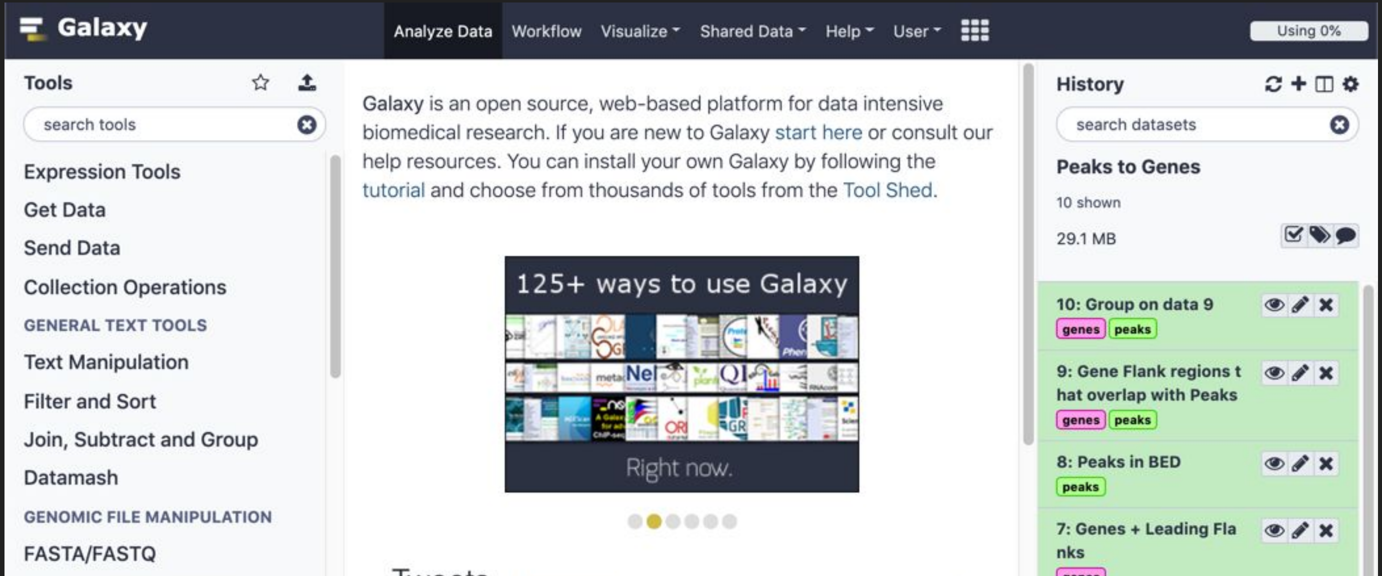
Dave Clements

Johns Hopkins University
clements@galaxyproject.org
@tnabtaf

Compute!

What is Galaxy?

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



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.

What is Galaxy?

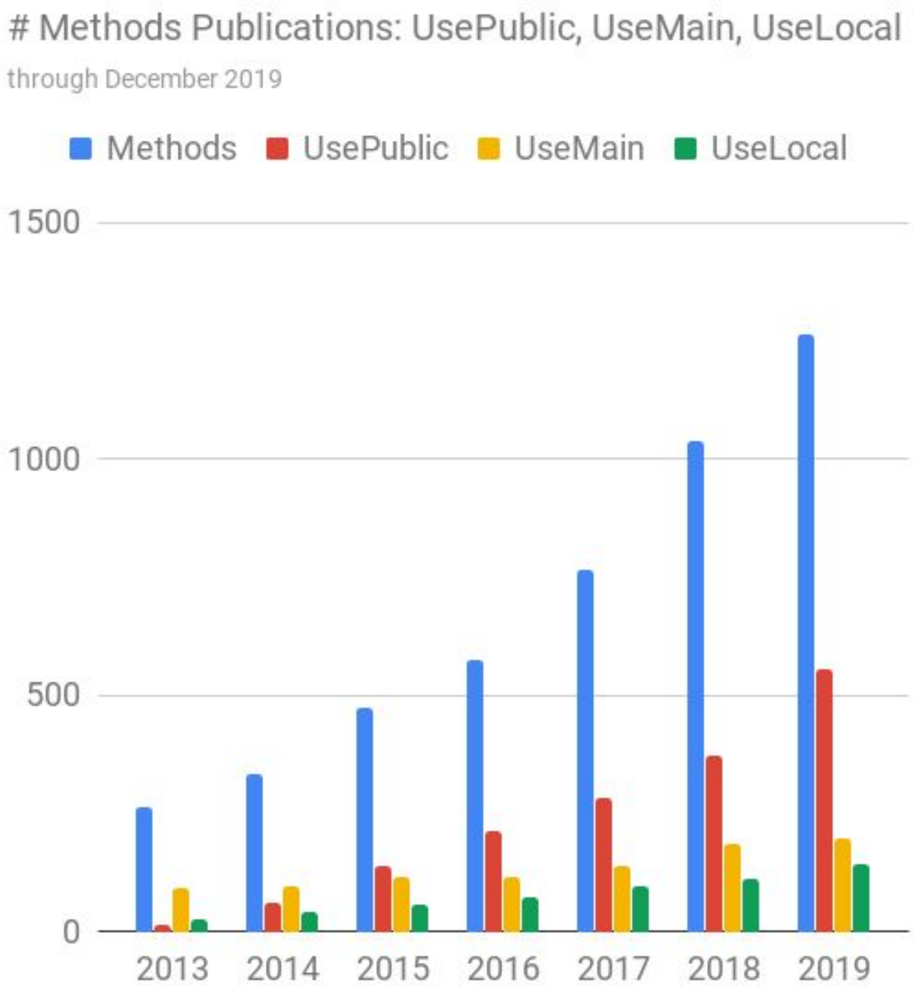
Users interact with data and tools via a graphical user interface. No computational experience required.

See Keith Bradnam's 13 Questions You May Have About Galaxy, 2015

Who Uses Galaxy?

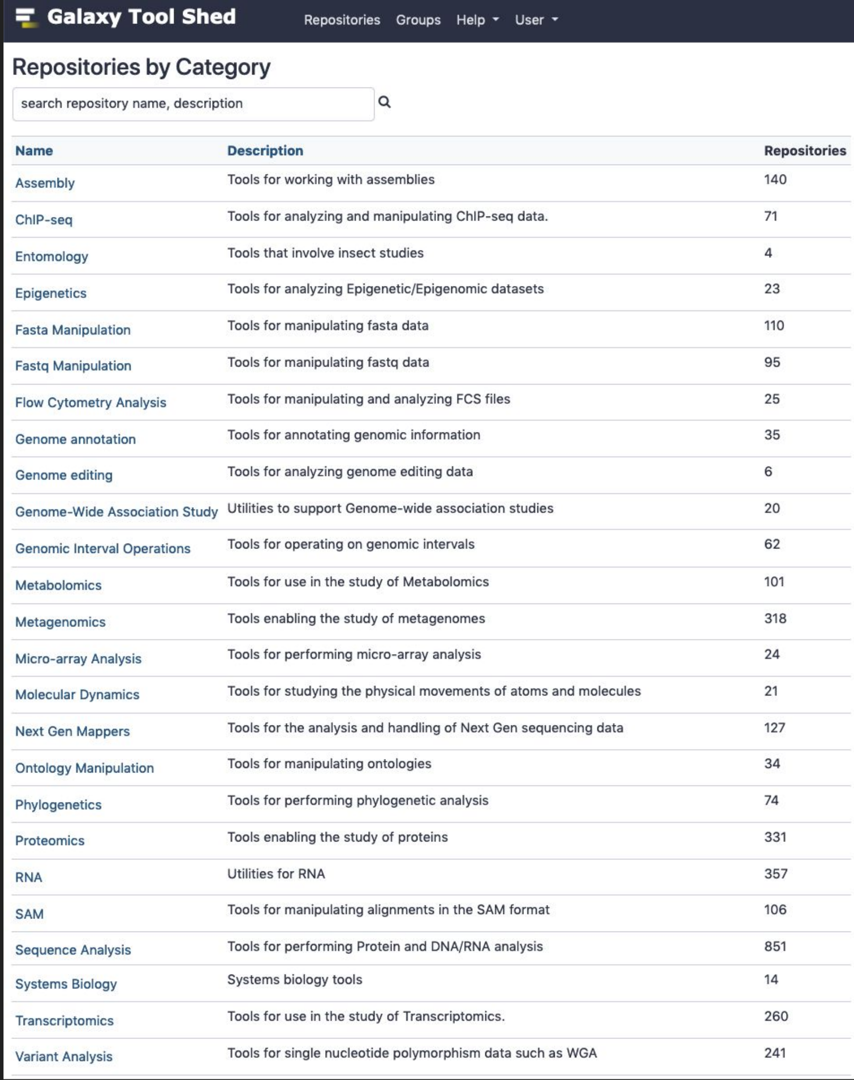
- 161 public platforms
- 100s (1000s?) of local installs
- Mentioned in almost 9000 pubs
  - in over 5000 methods sections
- UseGalaxy.org has over 170,000 registered users

galaxyproject.org/galaxy-project/statistics/



Who uses Galaxy: Omics

- Assembly
- ChIP-Seq & Epigenetics
- Flow Cytometry
- Genome Annotation
- Genome Editing
- GWAS
- Metabolomics
- Metagenomics
- Mapping
- Ontologies
- Phylogentics
- Proteomics
- RNA & Transcriptomics
- Sequence Analysis
- Systems Biology
- Variant Analysis



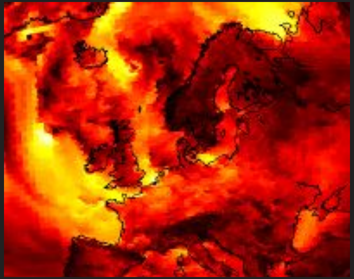
Omics Galaxy Toolshed Categories

Repositories by Category

Name	Description	Repositories
Assembly	Tools for working with assemblies	140
ChIP-seq	Tools for analyzing and manipulating ChIP-seq data.	71
Entomology	Tools that involve insect studies	4
Epigenetics	Tools for analyzing Epigenetic/Epigenomic datasets	23
Fasta Manipulation	Tools for manipulating fasta data	110
Fastq Manipulation	Tools for manipulating fastq data	95
Flow Cytometry Analysis	Tools for manipulating and analyzing FCS files	25
Genome annotation	Tools for annotating genomic information	35
Genome editing	Tools for analyzing genome editing data	6
Genome-Wide Association Study	Utilities to support Genome-wide association studies	20
Genomic Interval Operations	Tools for operating on genomic intervals	62
Metabolomics	Tools for use in the study of Metabolomics	101
Metagenomics	Tools enabling the study of metagenomes	318
Micro-array Analysis	Tools for performing micro-array analysis	24
Molecular Dynamics	Tools for studying the physical movements of atoms and molecules	21
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	127
Ontology Manipulation	Tools for manipulating ontologies	34
Phylogenetics	Tools for performing phylogenetic analysis	74
Proteomics	Tools enabling the study of proteins	331
RNA	Utilities for RNA	357
SAM	Tools for manipulating alignments in the SAM format	106
Sequence Analysis	Tools for performing Protein and DNA/RNA analysis	851
Systems Biology	Systems biology tools	14
Transcriptomics	Tools for use in the study of Transcriptomics.	260
Variant Analysis	Tools for single nucleotide polymorphism data such as WGA	241

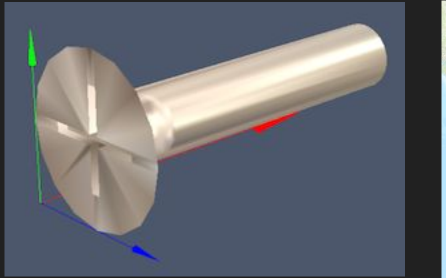
Who uses Galaxy: Other domains

Climate Science Workbench

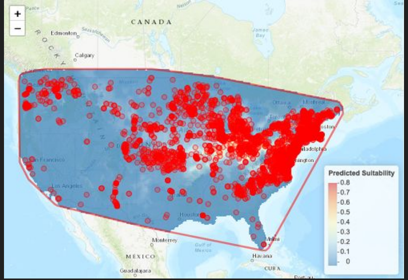


Complex Social Science Gateway

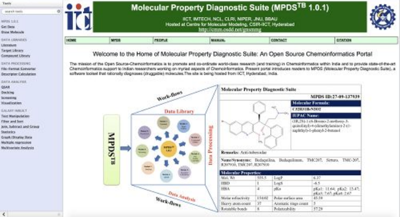
Galaxy for Constructive Solid Geometry



Galaxy for Ecology



Drug Development



Molecular Property Diagnostic Suite (MPDS^{TB} 1.0.1)

IICT, IMTECH, NCL, CLRI, NIPER, JNU, BBAU
Hosted at Centre for Molecular Modeling, CSIR-IICT, Hyderabad

Welcome to the Home of Molecular Property Diagnostic Suite: An Open Source Chemoinformatics Portal

The mission of the Open Source-Chemoinformatics is to promote and co-ordinate world-class research (and training) in Chemoinformatics within India and to provide state-of-the-art Chemoinformatics support to Indian researchers working on myriad aspects of Chemoinformatics. Present portal introduces readers to MPDS (Molecular Property Diagnostic Suite), a software toolset that rationally diagnoses (draggable) molecules.The site is being hosted from IICT, Hyderabad, India.

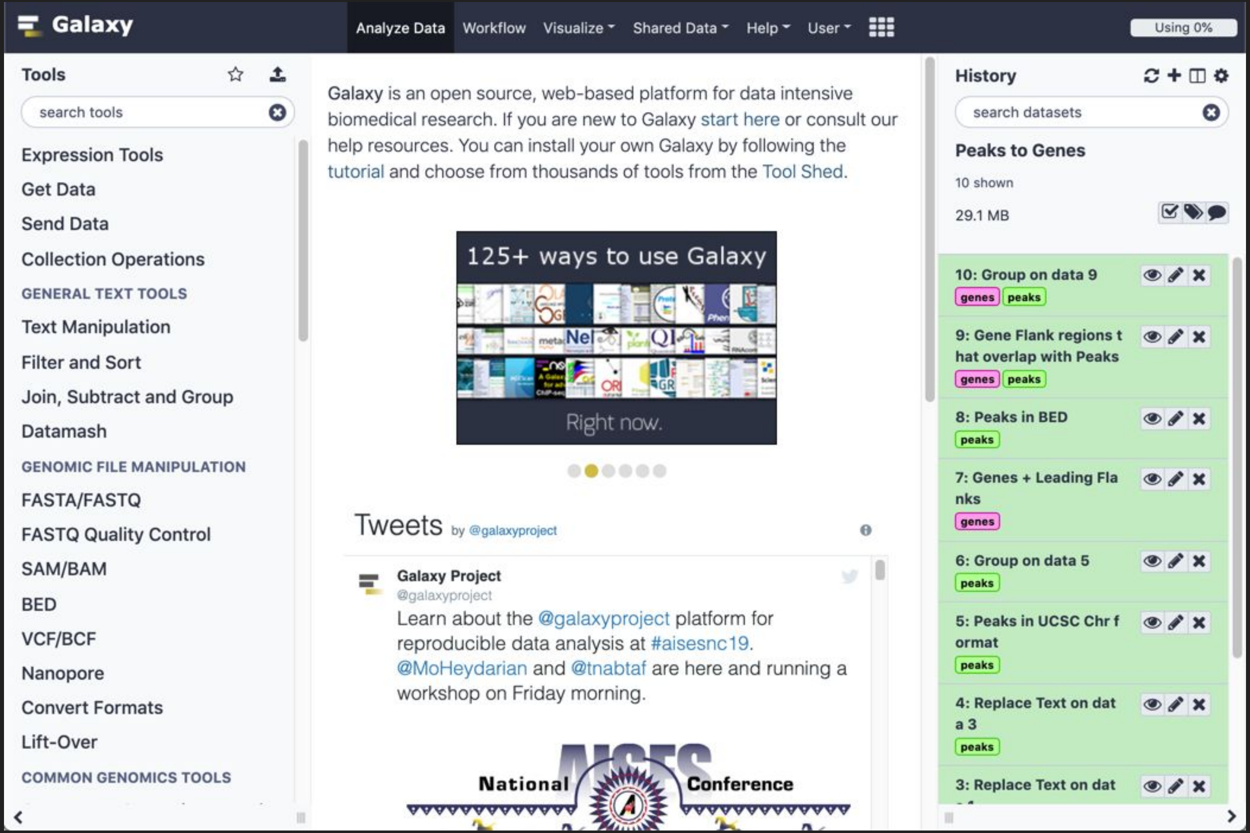
Molecular Property Diagnostic Suite

MPDS ID:27-09-137939

Natural Language Processing

CLARINO

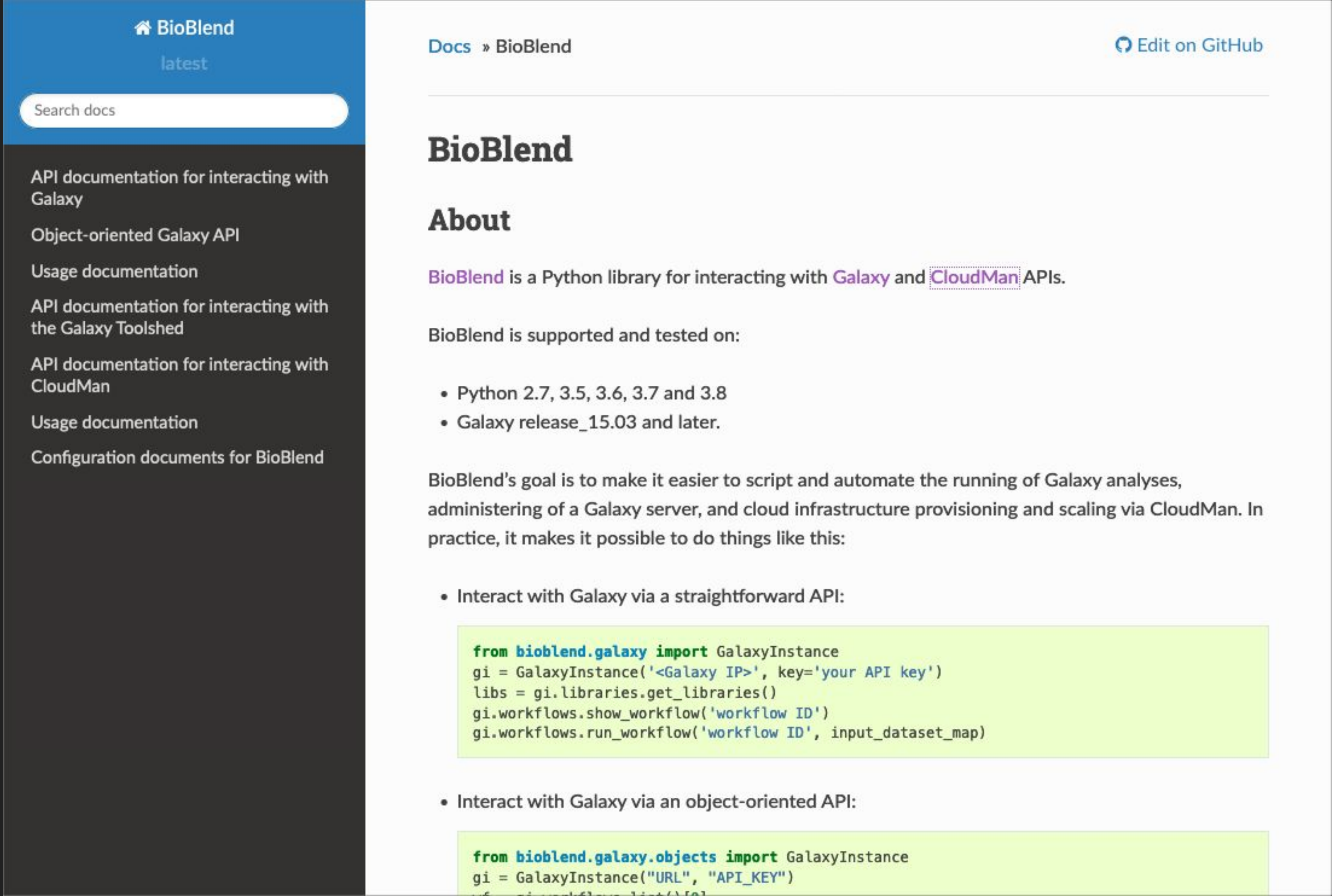
The Galaxy interface



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
Learn about the @galaxyproject platform for reproducible data analysis at #aisesnc19. @MoHeydarian and @tnabtaf are here and running a workshop on Friday morning.

The Galaxy Interface:
BioBlend API



BioBlend

About

BioBlend is a Python library for interacting with Galaxy and CloudMan APIs.

BioBlend is supported and tested on:

- Python 2.7, 3.5, 3.6, 3.7 and 3.8
- Galaxy release_15.03 and later.

BioBlend's goal is to make it easier to script and automate the running of Galaxy analyses, administering of a Galaxy server, and cloud infrastructure provisioning and scaling via CloudMan. In practice, it makes it possible to do things like this:

- Interact with Galaxy via a straightforward API:

```
from bioblend.galaxy import GalaxyInstance
gi = GalaxyInstance('<Galaxy IP>', key='your API key')
libs = gi.libraries.get_libraries()
gi.workflows.show_workflow('workflow ID')
gi.workflows.run_workflow('workflow ID', input_dataset_map)
```

- Interact with Galaxy via an object-oriented API:

```
from bioblend.galaxy.objects import GalaxyInstance
gi = GalaxyInstance("URL", "API_KEY")
```

Galaxy: Ecosystem & Community

How is Galaxy available?

Where	How soon	Choices
Public servers on the web	Right now	120+ web sites (UseGalaxy.*, RepeatExplorer, PhenoMeNal, Cistrome, Phylogeny.fr, ...)
Your own laptop	In a few minutes	30+ containers (Docker) and Virtual Machine images
On the cloud	In a few minutes to a few days	Academic (Jetstream, Nectar, CLIMB, GenAP, ...) and commercial (AWS, Google Cloud Platform, Azure) clouds
Your organization's own infrastructure	In a few weeks to a few months	100's (or 1000's) of local deployments

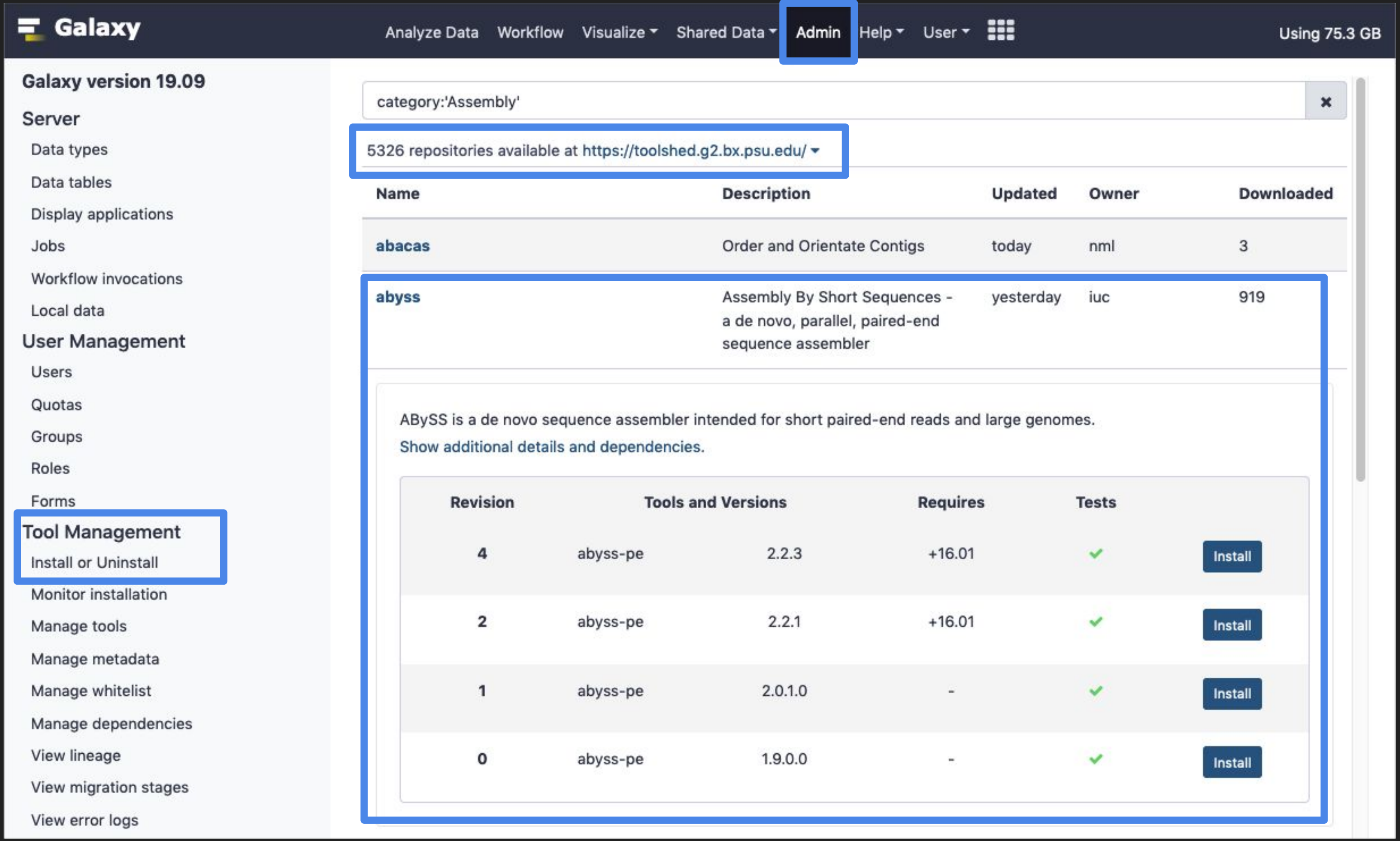
galaxyproject.org/use getgalaxy.org

Galaxies are Independent

- Galaxy instances are not interconnected (*yet*)
  - User identity, data, workflows are not connected between instances.
- Workflows can be exported from one instance and imported to others.
- Datasets can be exported from any instance

Tools: Galaxy Toolshed

1000s of tools & datatypes have been wrapped for Galaxy and are available for installation in servers though the Galaxy Admin GUI



category:'Assembly'

5326 repositories available at https://toolshed.g2.bx.psu.edu/

Name	Description	Updated	Owner	Downloaded
abacas	Order and Orientate Contigs	today	nml	3
abyss	Assembly By Short Sequences - a de novo, parallel, paired-end sequence assembler	yesterday	iuc	919

ABySS is a de novo sequence assembler intended for short paired-end reads and large genomes.
Show additional details and dependencies.

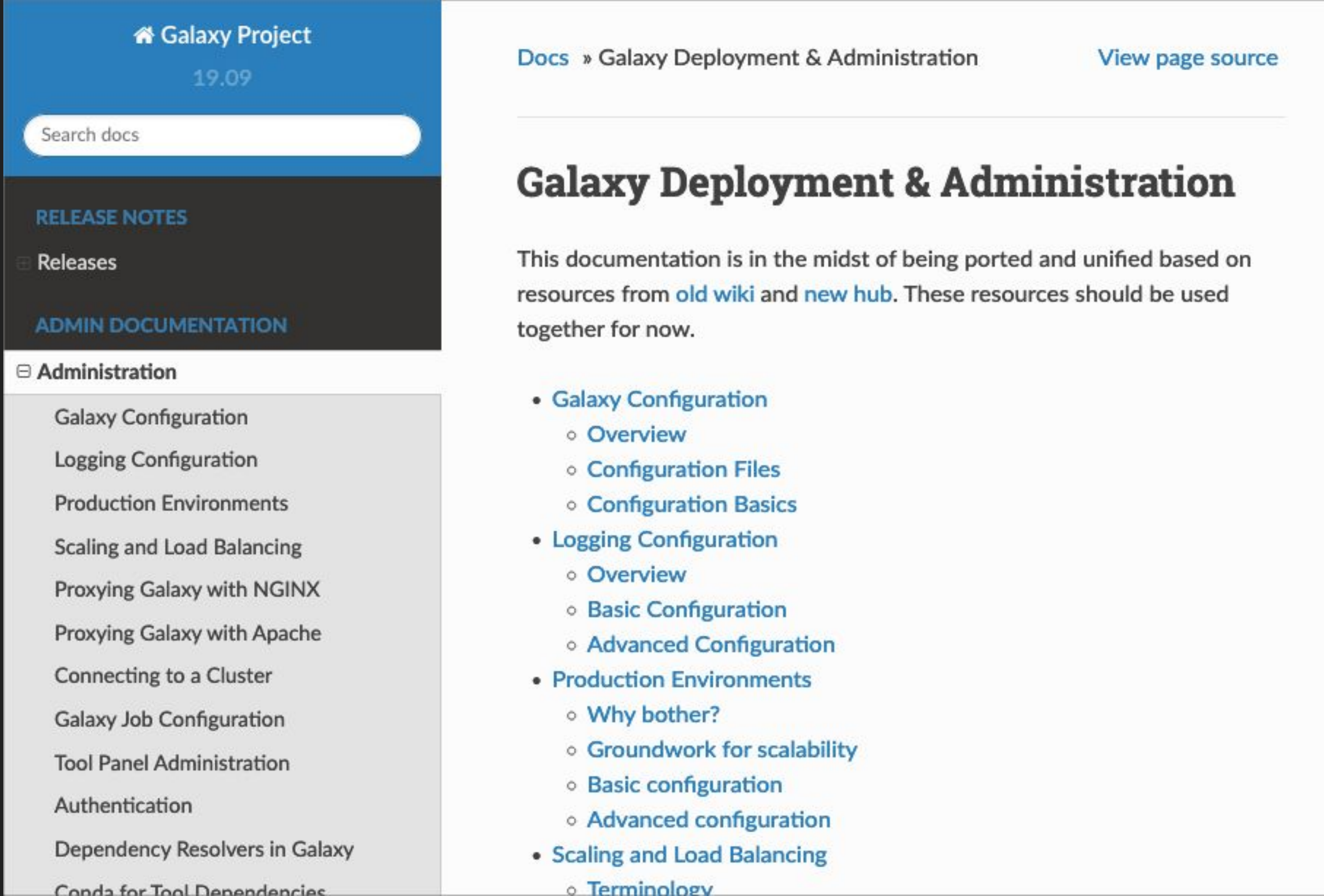
Revision	Tools and Versions		Requires	Tests	
4	abyss-pe	2.2.3	+16.01	✓	Install
2	abyss-pe	2.2.1	+16.01	✓	Install
1	abyss-pe	2.0.1.0	-	✓	Install
0	abyss-pe	1.9.0.0	-	✓	Install

Doc

Deployment & Admin

Using Galaxy

Tutorials ...



Galaxy Deployment & Administration

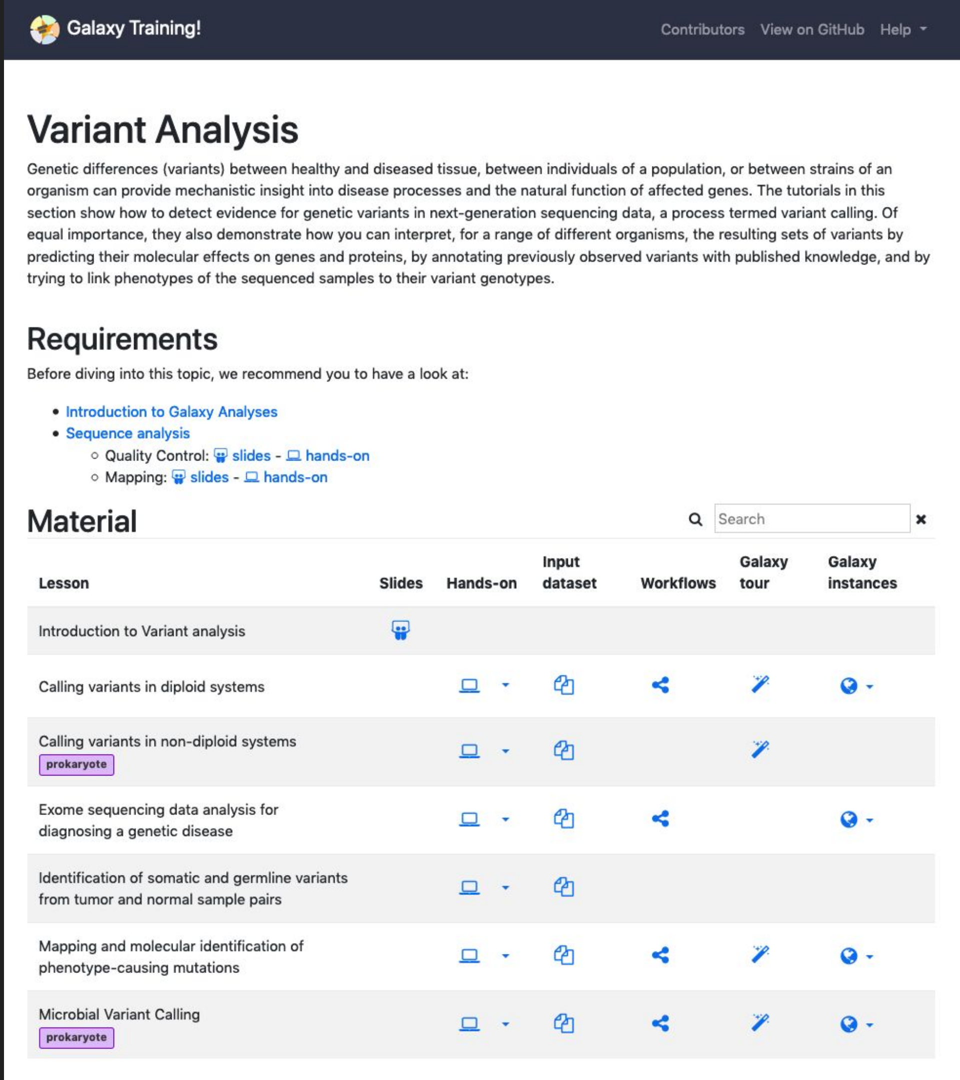
This documentation is in the midst of being ported and unified based on resources from old wiki and new hub. These resources should be used together for now.

- Galaxy Configuration
  - Overview
  - Configuration Files
  - Configuration Basics
- Logging Configuration
  - Overview
  - Basic Configuration
  - Advanced Configuration
- Production Environments
  - Why bother?
  - Groundwork for scalability
  - Basic configuration
  - Advanced configuration
- Scaling and Load Balancing
  - Terminology

Galaxy Training Network Library

- Slides
- Hands-on tutorials
- Training datasets
- Docker images
- Can be used individually or in classroom

training.galaxyproject.org/



Variant Analysis

Genetic differences (variants) between healthy and diseased tissue, between individuals of a population, or between strains of an organism can provide mechanistic insight into disease processes and the natural function of affected genes. The tutorials in this section show how to detect evidence for genetic variants in next-generation sequencing data, a process termed variant calling. Of equal importance, they also demonstrate how you can interpret, for a range of different organisms, the resulting sets of variants by predicting their molecular effects on genes and proteins, by annotating previously observed variants with published knowledge, and by trying to link phenotypes of the sequenced samples to their variant genotypes.

Requirements

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

- Introduction to Galaxy Analyses
- Sequence analysis
  - Quality Control: slides - hands-on
  - Mapping: slides - hands-on

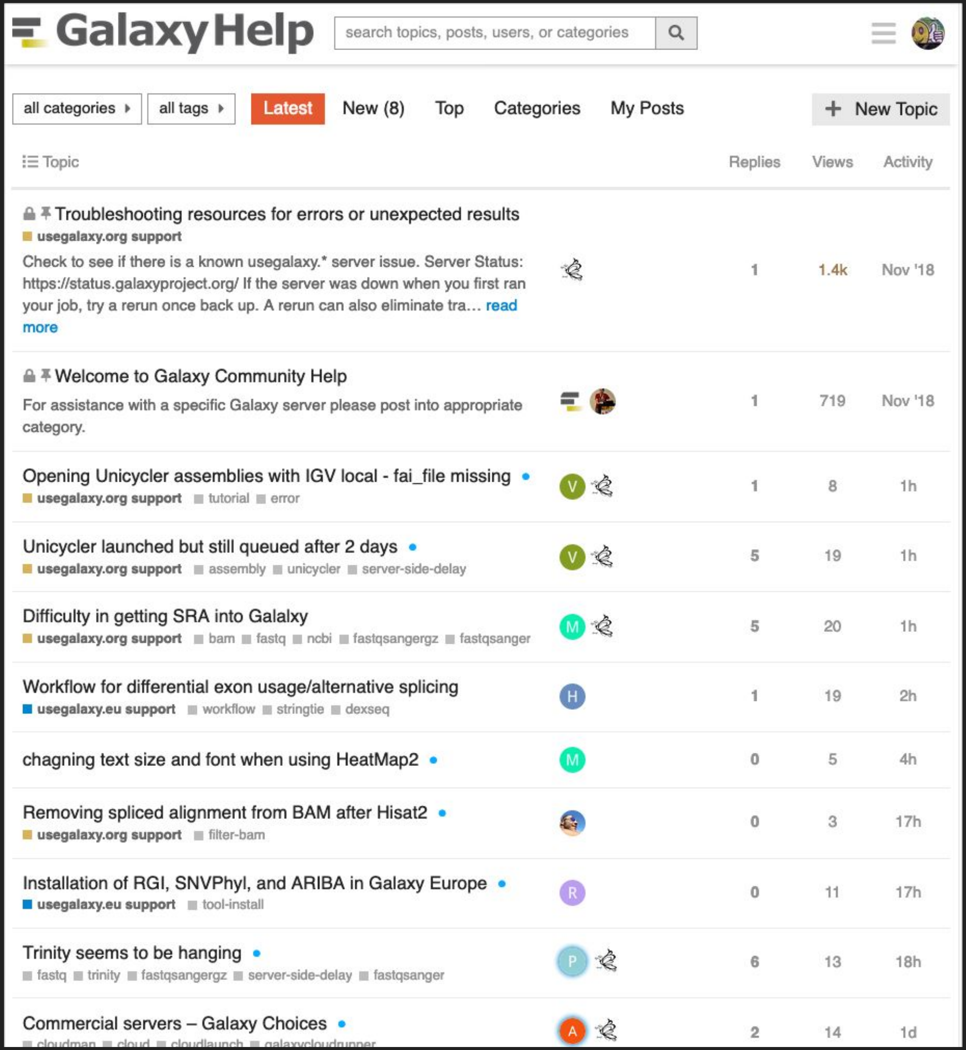
Material

Lesson	Slides	Hands-on	Input dataset	Workflows	Galaxy tour	Galaxy instances
Introduction to Variant analysis	✓					
Calling variants in diploid systems		✓	✓	✓	✓	✓
Calling variants in non-diploid systems prokaryote		✓	✓		✓	
Exome sequencing data analysis for diagnosing a genetic disease		✓	✓	✓		✓
Identification of somatic and germline variants from tumor and normal sample pairs		✓	✓			
Mapping and molecular identification of phenotype-causing mutations		✓	✓	✓	✓	✓
Microbial Variant Calling prokaryote		✓	✓	✓	✓	✓

Support

- Chat (Gitter channels)
- Online Forum (uses Discourse)
- Mailing Lists
- Doc
- Communities
- Videos

galaxyproject.org/support/



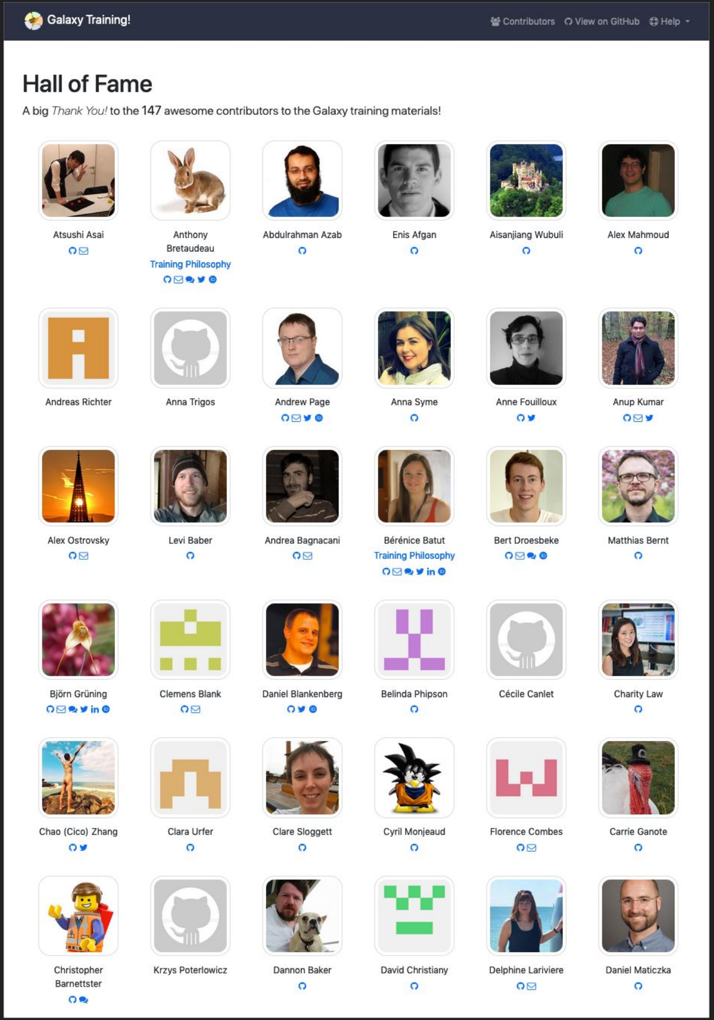
Topic	Replies	Views	Activity
Troubleshooting resources for errors or unexpected results usegalaxy.org support Check to see if there is a known usegalaxy.* server issue. Server Status: https://status.galaxyproject.org/ If the server was down when you first ran your job, try a rerun once back up. A rerun can also eliminate tra… read more	1	1.4k	Nov '18
Welcome to Galaxy Community Help For assistance with a specific Galaxy server please post into appropriate category.	1	719	Nov '18
Opening Unicycler assemblies with IGV local - fai_file missing usegalaxy.org support · tutorial · error	1	8	1h
Unicycler launched but still queued after 2 days usegalaxy.org support · assembly · unicycler · server-side-delay	5	19	1h
Difficulty in getting SRA into Galalxy usegalaxy.org support · bam · fastq · ncbi · fastqsangergz · fastqsanger	5	20	1h
Workflow for differential exon usage/alternative splicing usegalaxy.eu support · workflow · stringtie · dexseq	1	19	2h
chagning text size and font when using HeatMap2	0	5	4h
Removing spliced alignment from BAM after Hisat2 usegalaxy.org support · filter-bam	0	3	17h
Installation of RGI, SNVPhyl, and ARIBA in Galaxy Europe usegalaxy.eu support · tool-install	0	11	17h
Trinity seems to be hanging fastq · trinity · fastqsangergz · server-side-delay · fastqsanger	6	13	18h
Commercial servers – Galaxy Choices cloudman · cloud · cloudlaunch · galaxycloudrunner	2	14	1d

Contributors

Galaxy has an enormous and awesome contributor community

- 1,050 Help forum accounts in 13 months
- From BlackDuck Open Hub:
    - Over the past 12 months, 157 developers contributed new code to Galaxy … This is one of the largest open-source teams in the world *
- 147 contributors to GTN Library
- *ad infinitum*

* openhub.net/p/galaxybx/factoids



Hall of Fame

A big *Thank You!* to the 147 awesome contributors to the Galaxy training materials!

Atsushi Asai	Anthony Bretaudeau Training Philosophy	Abdulrahman Azab	Enis Afgan	Aisanjiang Wubuli	Alex Mahmoud
Andreas Richter	Anna Trigos	Andrew Page	Anna Syme	Anne Fouilloux	Anup Kumar
Alex Ostrovsky	Levi Baber	Andrea Bagnacani	Bérénice Batut Training Philosophy	Bert Droesbeke	Matthias Bernt
Björn Grüning	Clemens Blank	Daniel Blankenberg	Belinda Phipson	Cécile Canlet	Charity Law
Chao (Cico) Zhang	Clara Urfer	Clare Sloggett	Cyril Monjeaud	Florence Combes	Carrie Ganote
Christopher Barnettster	Krzys Poterlowicz	Dannon Baker	David Christiany	Delphine Lariviere	Daniel Maticzka

Events

- Galaxy & EIB @ PAG, January, San Diego
- Galaxy Admin Training, March, Barcelona
- BCC2020, July, Toronto
- Cornell, 2020 (working on it)
- Many other events, all over the world

galaxyproject.org/events/

Thank you

Bérénice Batut, GCC2019

Galaxy Community

The literally thousands of people who have contributed Tools, Doc, Support, Training, Resources, Code, Issue Reporting, Testing … over the past 15 years

Alexis Dereeper, Umesh Rosyara
Star Yanxin Gao, Mathias Lorieux, Ken McNally

PAG XXVIII

You