

Running Galaxy on the Cloud

Enis Afgan & Dannon Baker
Ruđer Bošković Institute & Emory University

@ GCC 2013, Oslo

Need an analysis? There's a tool for that.



Get Data

- [Upload File](#) from your computer
- [UCSC Main](#) table browser
- [UCSC Test](#) table browser
- [UCSC Archaea](#) table browser
- [BX main](#) browser
- [EBI SRA](#) ENA SRA
- [Get Microbial Data](#)
- [BioMart Central](#) server
- [BioMart Test](#) server
- [CBI Rice Mart](#) rice mart
- [GrameneMart](#) Central server
- [modENCODE fly](#) server
- [Flymine](#) server
- [Flymine test](#) server
- [modENCODE modMine](#) server
- [Ratmine](#) server
- [YeastMine](#) server
- [metabolicMine](#) server
- [modENCODE worm](#) server
- [WormBase](#) server
- [Wormbase](#) test server
- [EuPathDB](#) server
- [EncodeDB](#) at NHGRI
- [EpiGRAPH](#) server
- [EpiGRAPH](#) test server

NGS: Mapping

- [Lastz](#) map short reads against reference sequence
- [Lastz paired reads](#) map short paired reads against reference sequence
- [Map with Bowtie for Illumina](#)
- [Map with Bowtie for SOLiD](#)
- [Map with BWA for Illumina](#)
- [Map with BWA for SOLiD](#)
- [Map with BFAST](#)
- [Megablast](#) compare short reads against htgs, nt, and wgs databases
- [Parse blast XML output](#)
- [Map with PerM for SOLiD and Illumina](#)
- [Re-align with SRMA](#)
- [Map with Mosaik](#)

NGS: RNA Analysis

RNA-SEQ

- [Tophat for Illumina](#) Find splice junctions using RNA-seq data
- [Tophat for SOLiD](#) Find splice junctions using RNA-seq data
- [Cufflinks](#) transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- [Cuffcompare](#) compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- [Cuffdiff](#) find significant changes in transcript expression, splicing, and promoter use

FILTERING

- [Filter Combined Transcripts](#) using tracking file

NGS: GATK Tools (beta)

ALIGNMENT UTILITIES

- [Depth of Coverage](#) on BAM files

REALIGNMENT

- [Realigner Target Creator](#) for use in local realignment
- [Indel Realigner](#) - perform local realignment

BASE RECALIBRATION

- [Count Covariates](#) on BAM files
- [Table Recalibration](#) on BAM files
- [Analyze Covariates](#) - draw plots

GENOTYPING

- [Unified Genotyper](#) SNP and indel caller

ANNOTATION

- [Variant Annotator](#)

FILTRATION

- [Variant Filtration](#) on VCF files

VARIANT QUALITY SCORE RECALIBRATION

- [Variant Recalibrator](#)
- [Apply Variant Recalibration](#)

VARIANT UTILITIES

- [Validate Variants](#)
- [Eval Variants](#)
- [Combine Variants](#)

Tools

search tools

- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Convert Formats
- FASTA manipulation
- Filter and Sort
- Join, Subtract and Group
- 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
- Genome Diversity
- Phenotype Association
- EMBOSS
- NGS TOOLBOX BETA
- NGS: QC and manipulation
- NGS: Mapping
- NGS: SAM Tools
- NGS: GATK Tools (beta)
- ALIGNMENT UTILITIES
- Depth of Coverage on BAM files

Depth of Coverage (version 0.0.2)

Choose the source for the reference list:
 Locally cached

BAM files
 -I,--input_file <input_file>

BAM file 1
 BAM file:
 9: <https://galaxy-vic.genome.edu.au/datasets/a6aaf8a0aa187613/display?to>

Add new BAM file

Using reference genome:
 Human (Homo sapiens) (b37): hg_g1k_v37
 -R,--reference_sequence <reference_sequence>

RefSeq Rod:
 Selection is Optional
 -geneList,--calculateCoverageOverGenes <calculateCoverageOverGenes>

Partition type for depth of coverage:
 Select All Unselect All
 sample
 readgroup
 library
 -pt,--partitionType <partitionType>

Summary coverage thresholds
 -ct,--summaryCoverageThreshold <summaryCoverageThreshold>

Add new Summary coverage threshold

Output format:
 rtable
 --outputFormat <outputFormat>

Basic or Advanced GATK options:
 Basic

Basic or Advanced Analysis options:
 Basic

Execute

History

Random samples
 1.1 GB

- 14: [GenomeSpace Exporter on data 8](#)
- 13: [GenomeSpace import on Galaxy History Item: Generate pileup](#)
- 10: [Generate pileup on data 9: converted pileup](#)
- 9: [https://galaxy-vic.genome.edu.au/datasets/a6aaf8a0aa187613/display?to_ext=bam](#)
- 8: [Count on data 7](#)
- 7: [https://galaxy-vic.genome.edu.au/datasets/fdff468d39537a58/display?to_ext=tabular](#)
- 6: [FastQC http://static.vlsci.unimelb.edu.au/nathanh/NA12878.chr22_exome.BWA_mapped.chr22_filtered.bam.html](#)
- 5: [http://galaxy-vic.genome.edu.au/datasets/870093c60d62e4ae/display?to_ext=gtf](#)
- 4: [Generate pileup on data 3: converted pileup](#)
- 3: [http://static.vlsci.unimelb.edu.au/nathanh/NA12878.chr22_exome.BWA_mapped.chr22_filtered.bam](#)
- 2: [Map with Bowtie for Illumina on data 1](#)
- 1: [p1-c-b-1.fastq](#)

usegalaxy.org

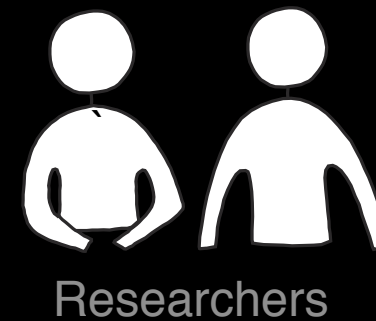
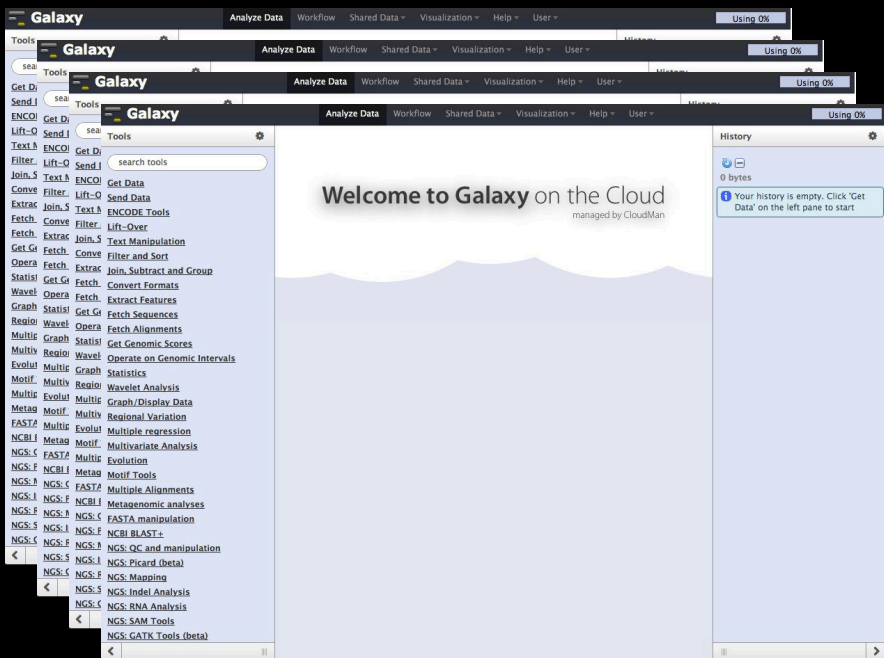
Getting your own Galaxy

Local install

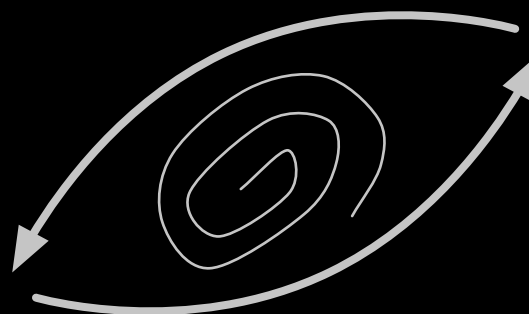
<http://wiki.galaxyproject.org/Admin/Get%20Galaxy>

Galaxy on the Cloud

<http://wiki.galaxyproject.org/CloudMan/>



Researchers



CloudMan

100sGB



100+

Cloud resources



usecloudman.org

When to use the cloud?

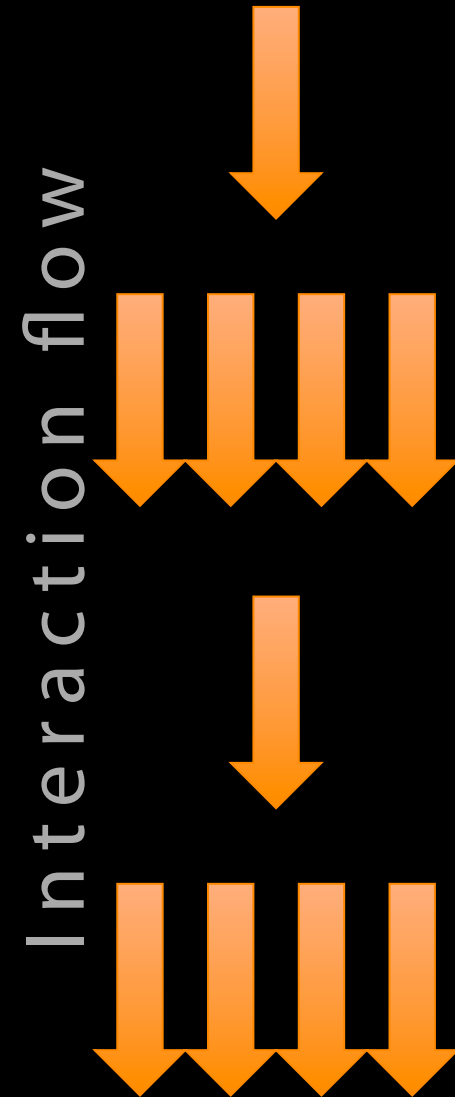
- Don't have informatics expertise or the infrastructure to run and maintain
- Have variable or particular resource needs
- Cannot upload data to a shared resource
- Need for customization
- Have oscillating data volume
- Want to test or share a tool, quickly & safely
- Want to make your analysis readily available to others
- Want fast access to AWS public datasets

How to use the cloud?

1. **Get an account** on the supported cloud
2. **Start a master instance** via the cloud web console or CloudLaunch
3. **Use CloudMan's web interface** on the master instance to manage the platform
4. **Use or customize Galaxy**

Workshop plan

- Launch an instance
- Demonstrate the following CloudMan features and prepare for the data analysis part:
 - Auto-scaling
 - Using an S3 bucket as a data source
 - Accessing an instance over ssh
 - Customizing an instance
 - Controlling Galaxy
 - Sharing-an-instance



YOUR TURN

Internet connection

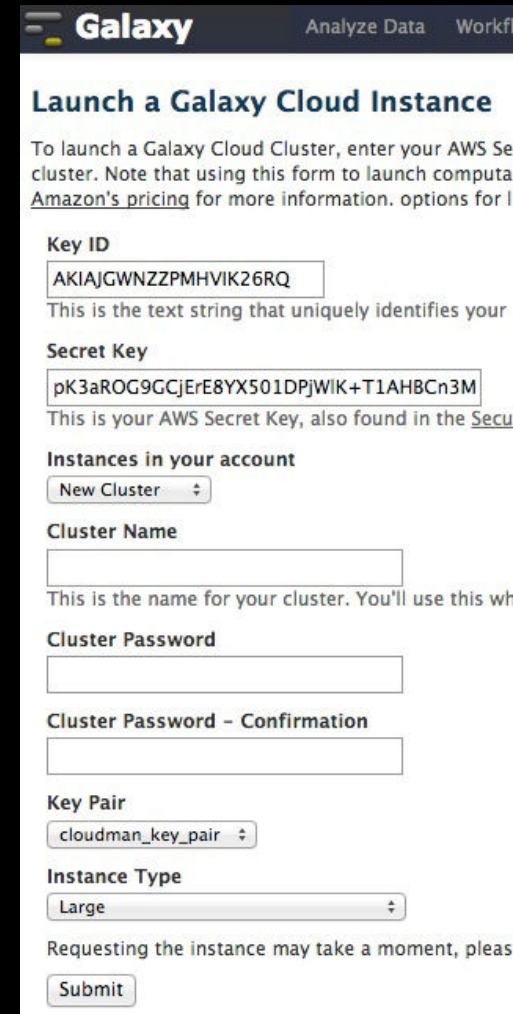
SSID: conferences

Password: uio202aar



Launch an instance

1. Visit usegalaxy.org/cloudlaunch
2. Enter the provided access key and secret key
Available from <http://bit.ly/gcw2013>
3. Choose New cluster
4. Set any name as the cluster name
5. Set any password
6. Choose CloudManKP1
7. Choose Large instance type
8. Launch your instance
Wait for the instance to start (~2-3 minutes)



The screenshot shows the 'Launch a Galaxy Cloud Instance' form in the Galaxy web interface. The form includes fields for Key ID (AKIAJGWNZZPMHVIK26RQ), Secret Key (pK3aROG9GCJErE8YX501DPJWIK+T1AHBCn3M), Cluster Name, Cluster Password, Cluster Password - Confirmation, Key Pair (cloudman_key_pair), and Instance Type (Large). A 'Submit' button is at the bottom.

Galaxy Analyze Data Workflows

Launch a Galaxy Cloud Instance

To launch a Galaxy Cloud Cluster, enter your AWS Security Credentials for an IAM user with access to the Amazon EC2 service. Note that using this form to launch compute instances is subject to [Amazon's pricing](#) for more information. options for I

Key ID

This is the text string that uniquely identifies your

Secret Key

This is your AWS Secret Key, also found in the [Secu](#)

Instances in your account

Cluster Name

This is the name for your cluster. You'll use this wh

Cluster Password

Cluster Password - Confirmation

Key Pair

Instance Type

Requesting the instance may take a moment, pleas

For more details, see
wiki.galaxyproject.org/CloudMan

Configure Your Cluster

CloudMan

Welcome to CloudMan. This application will allow you to manage this cluster and the services provided within. To get started, choose the type of cluster you'd like to work with and provide the associated value, if any.

Termination

Status

Cluster name

Disk status:

Worker status:

Service status:

Cluster status

Initial Cluster Configuration

Welcome to CloudMan. This application will allow you to manage this cluster and the services provided within. To get started, choose the type of cluster you'd like to work with and provide the associated value, if any.

Galaxy Cluster: Galaxy application, available tools, reference datasets, SGE job manager, and a data volume. Specify the initial storage size (in Gigabytes):

GB **OK** ← **20**

Share-an-Instance Cluster: derive your cluster from someone else's cluster. Specify the provided cluster share-string (for example, cm-0011923649e9271f17c4f83ba6846db0/shared/2011-08-19--21-00):

Cluster share-string

Data Cluster: a persistent data volume and SGE. Specify the initial storage size (in Gigabytes):

GB

Test Cluster: SGE only. No persistent storage is created.

[Hide extra options](#)

Choose CloudMan Platform

Manage Your Cluster

CloudMan Console

Welcome to [CloudMan](#). This application allows you to manage this instance cloud cluster and the services provided within. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to manage services provided by the application.

Terminate cluster

Add nodes ▼

Remove nodes ▼

Access Galaxy

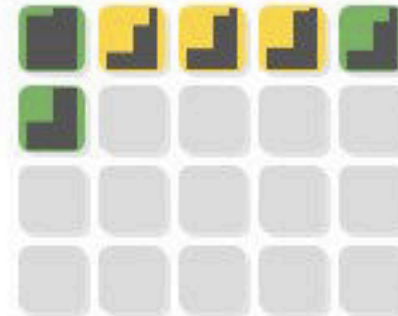
Status

Cluster name: ghem

Disk status: 0 / 0 (0%)

Worker status: Idle: 4 Available: 2 Requested: 5

Service status: Applications Data




Autoscaling is **off**.
Turn on?

Cluster status log

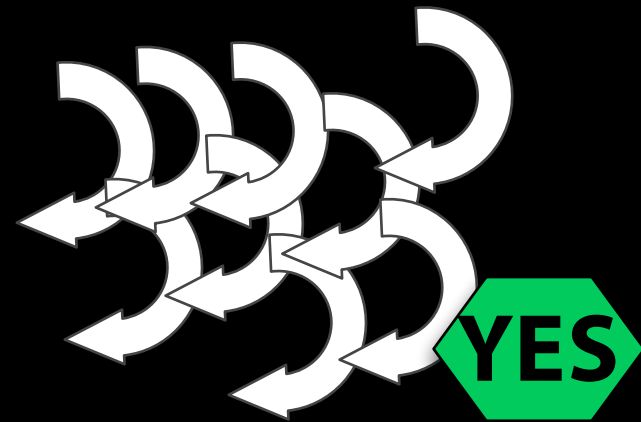
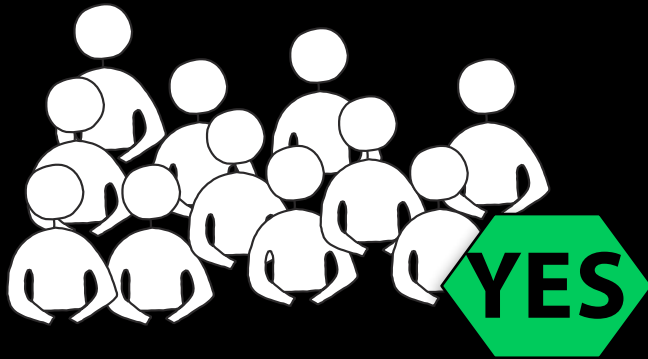


Scaling computation



i-79da2913
State: Ready
Alive: 1m 58s

- Filesystems
- Permissions
- Scheduler



Tools



search tools

[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](#)[Wavelet Analysis](#)[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](#)[NGS: SAM Tools](#)[NGS: GATK Tools \(beta\)](#)

Welcome to Galaxy on the Cloud

managed by CloudMan

Use your own Galaxy instance

- Register as a new user
- Become Admin user
- Use or customize

History

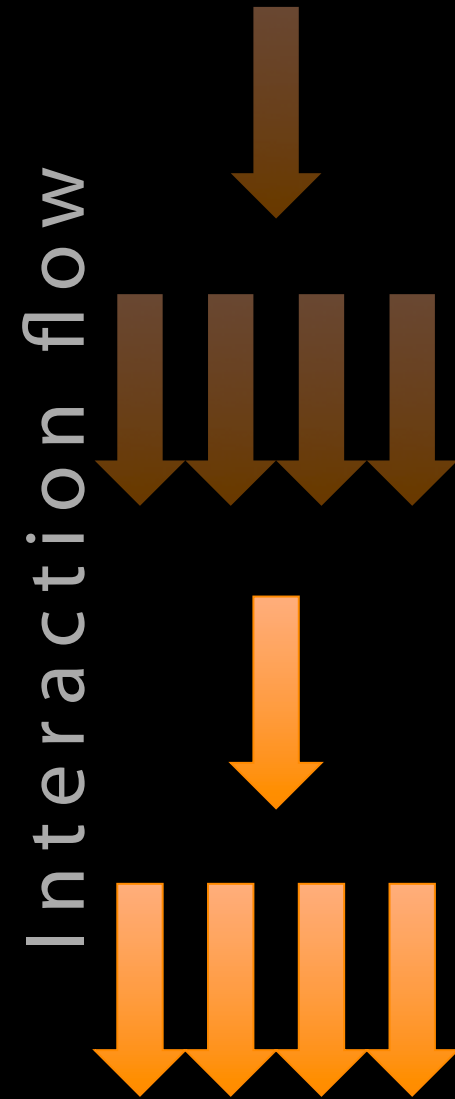


0 bytes

i Your history is empty. Click 'Get Data' on the left pane to start

Workshop plan

- Launch an instance ✓
- Demonstrate the following CloudMan features and prepare for the data analysis part:
 - Auto-scaling
 - Using an S3 bucket as a data source
 - Accessing an instance over ssh
 - Customizing an instance
 - Controlling Galaxy
 - Sharing-an-instance



Auto-scaling

Autoscaling is **off**.
Turn on?



Autoscaling Configuration

Autoscaling attempts to automate the elasticity offered by cloud computing for this particular cluster. **Once turned on, autoscaling takes over the control over the size of your cluster.**

Autoscaling is simple, just specify the cluster size limits you want to work within and use your cluster as you normally do. The cluster will not automatically shrink to less than the minimum number of worker nodes you specify and it will never grow larger than the maximum number of worker nodes you specify.

While respecting the set limits, if there are more jobs than the cluster can comfortably process at a given time autoscaling will automatically add compute nodes; if there are cluster nodes sitting idle at the end of an hour autoscaling will terminate those nodes reducing the size of the cluster and your cost.

Once turned on, the cluster size limits respected by autoscaling can be adjusted or autoscaling can be turned off.

Minimum number of nodes to maintain:

OK

0

Maximum number of nodes to maintain:

OK

3

Type of Node(s):

Same as Master

Turn autoscaling on

Auto-scaling in action

Fixed cluster size

5 nodes

Computation time: 9 hrs

Computation cost: \$20

20 nodes

Computation time: 6 hrs

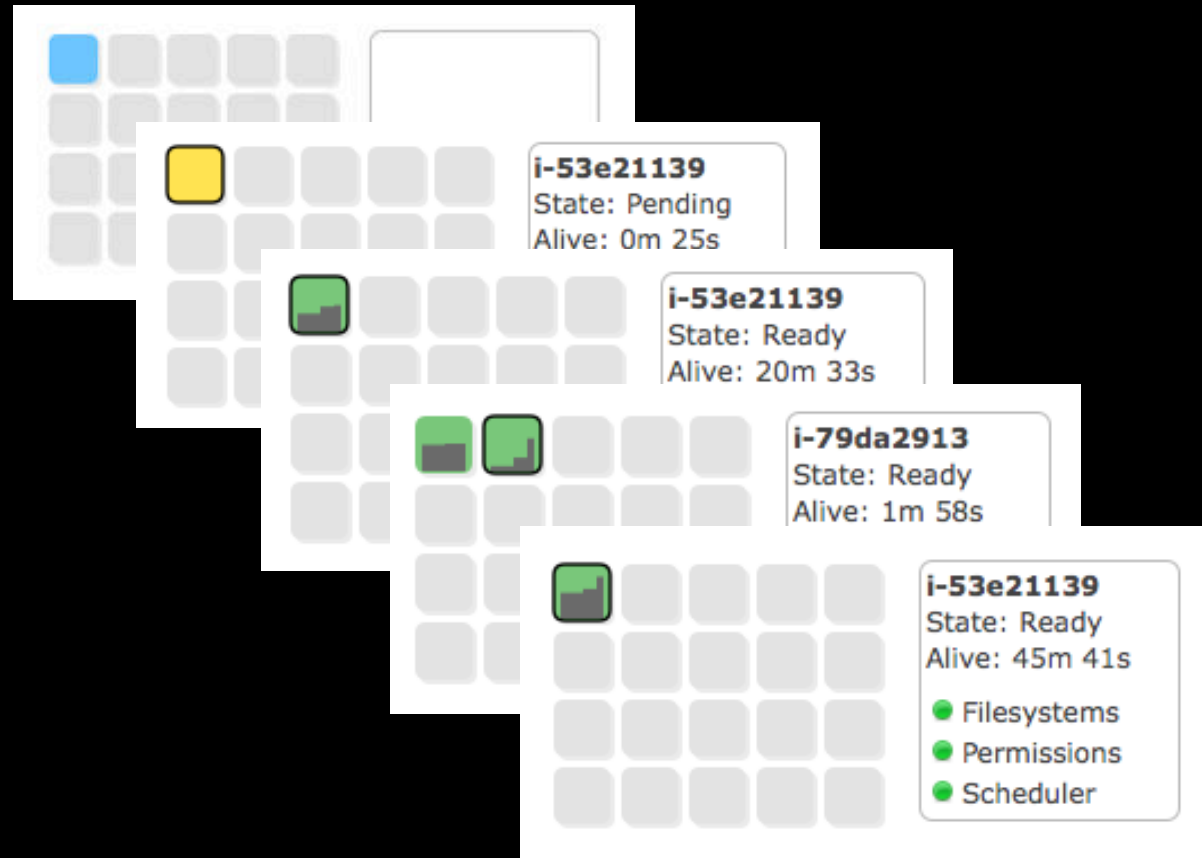
Computation cost: \$50

Dynamic cluster size

1 to 16
nodes

Computation time: 6 hrs

Computation cost: \$20



Using an S3 bucket as a data source

CloudMan from Galaxy [Admin](#) | [Report bugs](#) | [Wiki](#) | [Screencast](#)

Available file systems

Add a new file system

This form allows you to add an additional data source and make it available as a local file system. Currently, adding S3 buckets as a data source is the only supported functionality. These buckets may be public or private (and owned by the user running this cluster). Once added, the file system will be available on the underlying system under `/mnt/[bucket_name]` path.

Bucket name: **3: workshop-data**

- Running Galaxy at revision: [8140:8afb981f46c1](#)
- Update Galaxy from a provided repository What will this do?

Services controls

Use these controls to administer individual application services managed by CloudMan. Currently running a 'Galaxy' type of cluster.

Name	Status	Log	Stop	Start	Restart	Update DB
Galaxy	Running	Log	Stop	Start	Restart	Update DB
PostgreSQL	Running	Log	Stop	Start	Restart	
SGE	Running	Log	Stop	Start	Restart	Q conf gstat
File systems	Running	No logs	Manage			

System controls

Use these controls to administer CloudMan itself as well as the underlying system.

1 (points to Admin link)

2 (points to Stop button in Services controls table)

Accessing an instance over ssh

Install *Secure Shell from Chrome*

SSH using user ubuntu and the password you chose when launching an instance

Once logged in

- You have full system access to your instance, including sudo; use it as any other system
- *galaxy* user exists on the system and should be used when manipulating Galaxy (sudo su galaxy)
- Can submit any jobs via the standard qsub command

Customizing an instance

- Edit Galaxy's configuration

```
$ sudo su galaxy
```

```
$ cd /mnt/galaxy/galaxy-app
```

```
$ vi universe_wsgi.ini
```

```
allow_library_path_paste = True
```


Controlling Galaxy

- Start/stop Galaxy application
- Add an admin user
 - Use the email you registered with

S3 bucket as a data library

- Within Galaxy, create a Data Library, using S3 bucket path as the data source (/mnt/workshop-data)
- This will import all the datasets into the Data Library
- Import that dataset into a history

Sharing-an-Instance

- Share the entire Galaxy CloudMan platform
 - Includes all of user data and even the customizations
- Publish a self-contained analysis
- Make a note of the *share-string* and send it to your neighbor

Currently shared instances

Share-an-instance

This form allows you to share this cluster instance, at its current state, with others. You can make the instance public or share it with specific users by providing their account information below. You may also share the instance with yourself by specifying your own credentials, which will have the effect of saving the instance at its current state.

While setting up an instance to be shared, all currently running cluster services will be stopped. Then, a snapshot of your data volume and a folder in your cluster's bucket will be created (under 'shared/[current date and time]'); this folder will contain your cluster's current configuration. The created snapshot and the folder will be given READ permissions to the users you choose (or make it public). This will enable those users to instantiate their own instances of the given cluster instance. This implies that you will only be paying for the created snapshot while users deriving a cluster from yours will incur costs for running the actual cluster. After the sharing process is complete, services on your cluster will automatically resume.

Public Shared **Public**

Specific user permissions:

Both fields must be provided for each of the users. These numbers can be obtained from the bottom of the AWS Security Credentials page, under Account Identifiers section.

AWS account numbers: CIV numbers

AWS canonical user IDs: CIV numbers

Share-an-instance

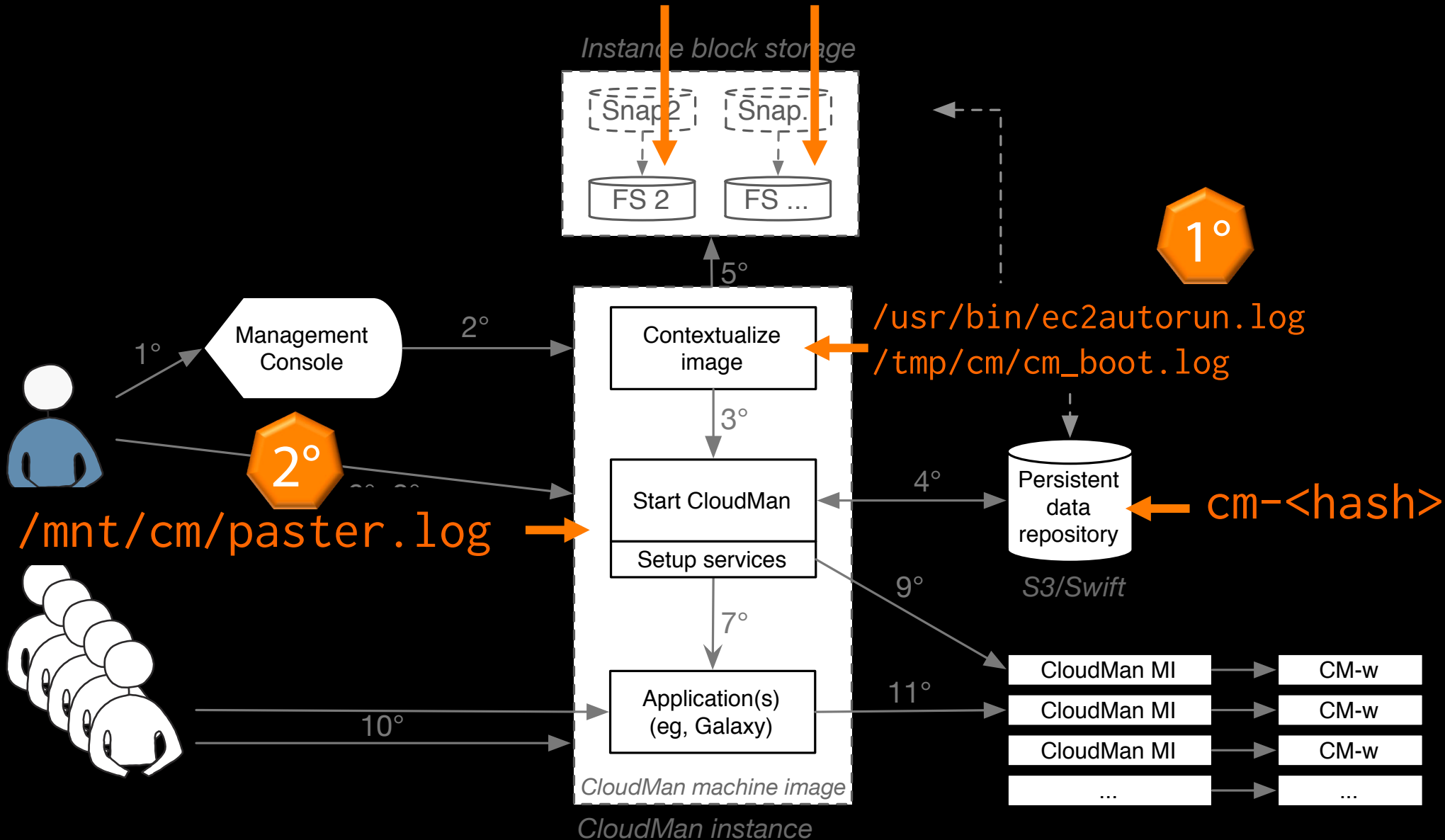
Name	Instance ID	References
Exome sequencing pipeline	cm-b53c6f1223f966914df347687f6fc818/shared/2011-10-07--14-00	Pipeline description

CloudMan Features

- Start/launch/control through a web browser
- API (via BioBlend library)
- Choose between multiple cluster types
- Terminate and recreate or restart
- Scale the size of the compute cluster
 - Auto-scale
- Support for AWS Spot instances
- Expand the file system
- Customize (via CLI or the Tool Shed): tools, data, references
- Share-an-instance (customized one too)
- Mount an S3 bucket -> data library
- Access via ssh
- Control the Galaxy process
- Deploy on your cloud (automated via CloudBioLinux)

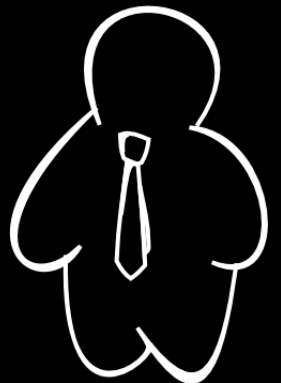
Troubleshooting

`/mnt/galaxy[Indices]`



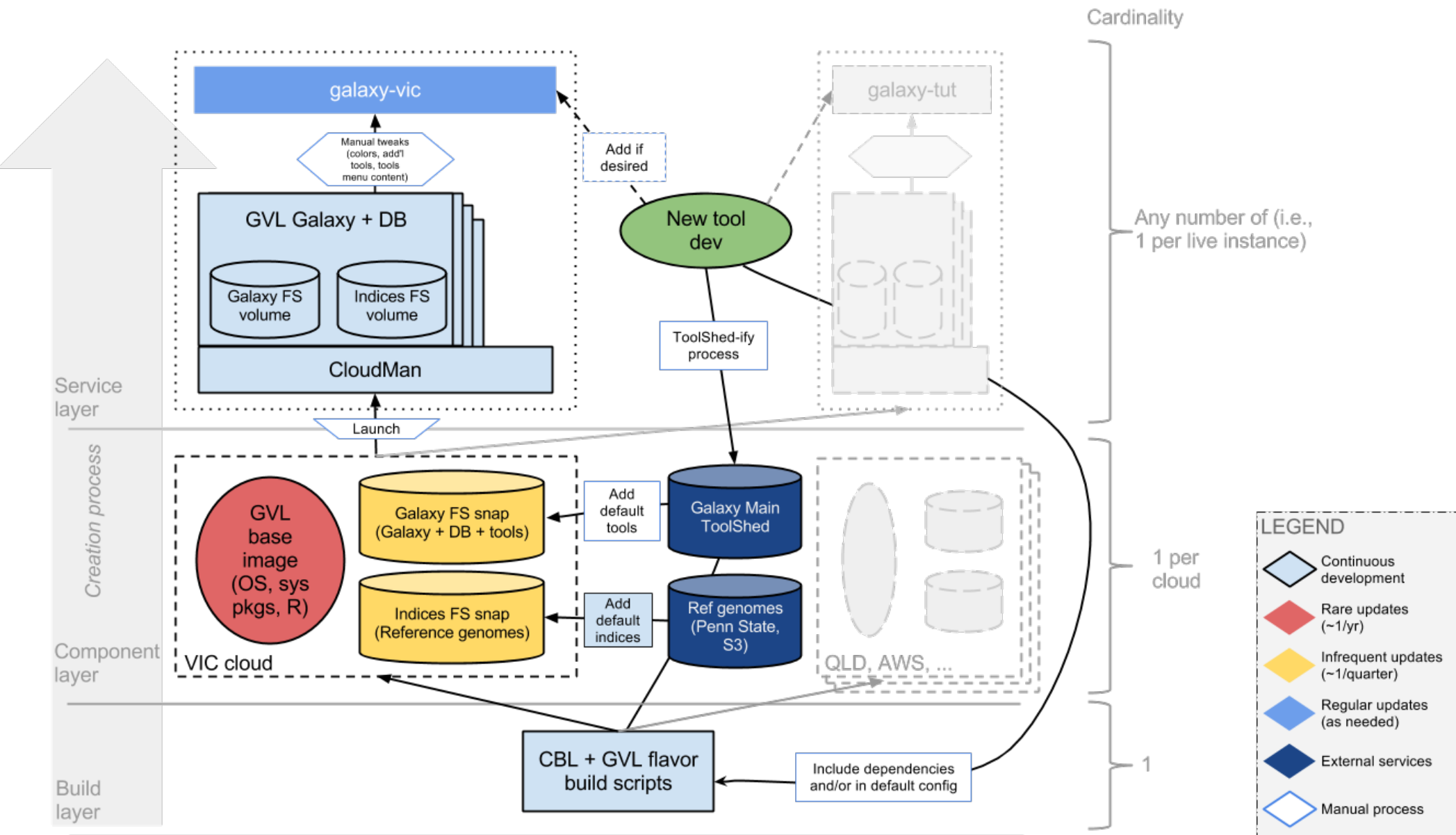
Running Galaxy CloudMan on **your own cloud**

A brief overview



Supported clouds

- Amazon Web Services
- OpenStack
- Eucalyptus
- OpenNebula



Building

- Leverage **CloudBioLinux** build framework
- Developed GVL flavor
 - Base Galaxy image
 - Full CloudBioLinux image
- There are also more specific flavors available
 - cloudman

github.com/afgane/gvl_flavor

Deploying

- Integrated with [BioCloudCentral.org](https://biocloudcentral.org)
 - Use the public one, deploy your own or run locally
- Supports multiple clouds
 - NeCTAR cloud
 - AWS Sydney

Contribute

(start with documentation)





Enis Afgan



Guru Ananda



Dannon Baker



Dan Blankenberg



Dave Bouvier



Dave Clements



Nate Coraor



Carl Eberhard



Jeremy Goecks



Nuwan
Goonasekera



Jen Jackson



Greg von Kuster



Ross Lazarus



Rémi Marengo



Scott McManus



Anton
Nekrutenko



James
Taylor

The Galaxy Team

<http://galaxyproject.org/wiki/GalaxyTeam>