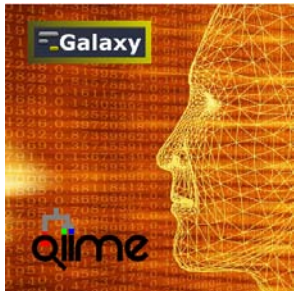


The Cloud for Biologists

using bioinformatics tools



Mattias de Hollander

Netherlands Institute of
Ecology (NIOO-KNAW)



Why choose for the Cloud?



Why choose for the Cloud?



- It's **flexible**



Why choose for the Cloud?



- It's **flexible**
- You have full **control**



Why choose for the Cloud?



- It's **flexible**
- You have full **control**
- Perfect for **small labs**



Why choose for the Cloud?



- It's **flexible**
- You have full **control**
- Perfect for **small labs**
- It's **fancy** (Google and Amazon are using it)



Why choose for the Cloud?



- It's **flexible**
- You have full **control**
- Perfect for **small labs**
- It's **fancy** (Google and Amazon are using it)
- It's **environmental friendly** (Gmail: Its cooler in the cloud)





How do we use the Cloud?



Galaxy

a web-based genome analysis platform¹



¹Slide by Anton Nekrutenko, Galaxy Developer Conference 2011, Lunteren (NL)



Galaxy

a web-based genome analysis platform¹



- A free (for everyone) **web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage

¹Slide by Anton Nekrutenko, Galaxy Developer Conference 2011, Lunteren (NL)



Galaxy

a web-based genome analysis platform¹



- A free (for everyone) **web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- **Open source software** that makes integrating your own tools and data and customizing for your own site simple

¹Slide by Anton Nekrutenko, Galaxy Developer Conference 2011, Lunteren (NL)



Galaxy
Analyze Data Workflow Shared Data Visualization Help User
Using 0%

Tools Options ▾

[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](#)

[Human Genome Variation](#)

[Genome Diversity](#)

[EMBOSS](#)

NGS TOOLBOX BETA

[NGS: QC and manipulation](#)

[NGS: Mapping](#)


[NGS: SAM Tools](#)

[NGS: Indel Analysis](#)

[NGS: Peak Calling](#)

[NGS: RNA Analysis](#)

Check out the new Galaxy Tool Shed



Live Quickies

Basic fastQ manipulation:

Galaxy quickie # 11

Advanced fastQ manipulation:

Galaxy quickie # 14

454 Mapping: Single End


Galaxy quickie # 11

Uploading Data using FTP

Galaxy quickie # 17

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The **Galaxy team** is a part of **ix** at Penn State, and the **Biology and Mathematics and Computer Science** departments at **Emory University**. 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 Emory University**.

Galaxy build: **9Rev 5957-5d2a2ac4710fs**

 galaxyproject

ChIP-Seq: technical considerations for obtaining high-quality data, Kibler et al., Nature Immunology, <http://t.co/WHh9ky>
16 hours ago · reply · retweet · favorite

"Galaxy Provides Life Support for NGS Exploration," article by Kevin Davies in Bio-ITWorld, http://t.co/QZdHm4E_Auegalaxy

History Options ▾

Unnamed history 117.5 KB

- 35: Extract Pairwise MAP blocks on data 23
- 34: Extract Pairwise MAP blocks on data 23
- 33: Extract Pairwise MAP blocks on data 23
- 32: Extract Pairwise MAP blocks on data 23
- 29: Extract Genomic DNA on data 28
- 28: UCSC Main on Human: knownGene (chr18:9092712-9124339)
- 27: Extract Genomic DNA on NDUFV2P1
- 26: Extract Genomic DNA on NDUFV2
- 25: NDUFV2
- 23: NDUFV2P1
- 16: Extract MAP blocks on data 9
- 15: Extract MAP blocks on data 9
- 11: Extract MAP blocks on data 9

Most biologists don't write code



Most biologists don't write code



- Analyze
 - Interactively **manipulate** genomic data with a comprehensive and expanding 'best-practices' toolset



Most biologists don't write code



- Analyze
 - Interactively **manipulate** genomic data with a comprehensive and expanding 'best-practices' toolset
- Publish and Share
 - Results and step-by-step analysis record (**Data Libraries** and **Histories**)
 - Customizable pipelines (**Workflows**)
 - Share workflows with other users





Cloudman



What is Cloudman?



What is Cloudman?



- Cloudman is written by Enis Afghani *et.al.*, Emory University and provides a ready-to-run, dynamically scalable version of Galaxy on **Amazon AWS**



What is Cloudman?



- Cloudman is written by Enis Afghan *et.al.*, Emory University and provides a ready-to-run, dynamically scalable version of Galaxy on **Amazon AWS**
- Now it's possible to run it also on the **SARA HPC Cloud / Opennebula** (with some limitations)



How does it work?



How does it work?



- A **master node** contains all the data and tools



How does it work?



- A **master node** contains all the data and tools
- Initiate **worker nodes** based on needs/load



How does it work?



- A **master node** contains all the data and tools
- Initiate **worker nodes** based on needs/load
- Data is available on all nodes using a **shared filesystem**: NFS



How does it work?



- A **master node** contains all the data and tools
- Initiate **worker nodes** based on needs/load
- Data is available on all nodes using a **shared filesystem**: NFS
- RabbitMQ is used for **communication** between cluster nodes



How does it work?



- A **master node** contains all the data and tools
- Initiate **worker nodes** based on needs/load
- Data is available on all nodes using a **shared filesystem**: NFS
- RabbitMQ is used for **communication** between cluster nodes
- Jobs are queued using **SGE**



How does it work?



- A **master node** contains all the data and tools
- Initiate **worker nodes** based on needs/load
- Data is available on all nodes using a **shared filesystem**: NFS
- RabbitMQ is used for **communication** between cluster nodes
- Jobs are queued using **SGE**
- Galaxy is served using nginx **webserver**



Workers instances are being configured



Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application allows you to manage this instance of Galaxy CloudMan. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to add and remove 'worker' nodes for running jobs.

[Terminate cluster](#)[Add nodes ▼](#)[Remove nodes ▼](#)[Access Galaxy](#)

Status

Cluster name: local test

Disk status: 0 / 0 (0%)

Worker status: Idle: 0 Available: 0 Requested: 3

Service status: Applications ● Data ●

External Logs: [Galaxy Log](#)



Autoscaling is **off**.
Turn on?

[Cluster status log](#)



Galaxy is accessible



The screenshot displays the Galaxy web interface. At the top, a dark navigation bar contains the 'Galaxy' logo and several menu items: 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. On the left side, there is a 'Tools' panel with an 'Options' dropdown and a list of tool categories such as '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', 'Graph/Display Data', 'Regional Variation', 'Multiple regression', 'Multivariate Analysis', 'Evolution', 'Motif Tools', 'Multiple Alignments', 'Metagenomic analyses', 'FASTA manipulation', 'NCBI BLAST+', 'NGS: QC and manipulation', and 'NGS: Picard'. The main content area features a large, light blue background with a white mountain range silhouette and the text 'Welcome to Galaxy on the Cloud'. A mouse cursor is visible over the background. On the right side, there is a 'History' panel with an 'Options' dropdown, showing '0 bytes' and a message: 'Your history is empty. Click 'Get Data' on the left pane to start.'





How is Galaxy used at the NIOO?



How is Galaxy used at the NIOO?



How is Galaxy used at the NIOO?



- Analyzing high-throughput community sequencing data with QIIME



How is Galaxy used at the NIOO?



- Analyzing high-throughput community sequencing data with QIIME
 - Denoising (**CPU-intensive**)



How is Galaxy used at the NIOO?



- Analyzing high-throughput community sequencing data with QIIME
 - Denoising (**CPU-intensive**)
 - OTU and representative set picking using uclust, cdhit, mothur BLAST or other tools



How is Galaxy used at the NIOO?



- Analyzing high-throughput community sequencing data with QIIME
 - Denoising (**CPU-intensive**)
 - OTU and representative set picking using uclust, cdhit, mothur BLAST or other tools
 - Taxonomy assignment with BLAST or the RDP classifier (**CPU-intensive**)



How is Galaxy used at the NIOO?



- Analyzing high-throughput community sequencing data with QIIME
 - Denoising (**CPU-intensive**)
 - OTU and representative set picking using uclust, cdhit, mothur BLAST or other tools
 - Taxonomy assignment with BLAST or the RDP classifier (**CPU-intensive**)
 - Sequence alignment with PyNAST, muscle, infernal, or other tools (**CPU-intensive**)



How is Galaxy used at the NIOO?



- Analyzing high-throughput community sequencing data with QIIME
 - Denoising (**CPU-intensive**)
 - OTU and representative set picking using uclust, cdhit, mothur BLAST or other tools
 - Taxonomy assignment with BLAST or the RDP classifier (**CPU-intensive**)
 - Sequence alignment with PyNAST, muscle, infernal, or other tools (**CPU-intensive**)
 - and more!





Thanks!



Thanks to the Galaxy Cloud Team



Enis Afgan



Dannon Baker



Dan Blankenberg



Nate Coraor



Dave Clements



Jeremy Goecks



Jennifer Jackson



Greg von Kuster



Kanwei Li



James Taylor



Kelly Vincent



Anton Nekrutenko



Questions?





Extra slides



Limitations of Opennebula



- Create instances providing user data (available in production cloud?)
- No support for growing qcow filesystem
- Would be create to access the cloud the ON API from outside
- Cloned instances have not a working network



More info at



- My notes:
<https://www.cloud.sara.nl/projects/galaxy/wiki>
- Galaxy Cloud on Amazon: <http://usegalaxy.org/cloud>
- Cloudman scripts:
<https://bitbucket.org/galaxy/cloudman/>
- Install tools:
<https://bitbucket.org/afgane/mi-deployment>
- Bio-linux repository: <http://nebc.nerc.ac.uk/tools/bio-linux/bio-linux-6.0>



Launch Cloudman Console



Galaxy Cloudman

Info: [report bugs](#) | [wiki](#) | [screencast](#)

Galaxy Cloudman

Welcome to Galaxy Cloudman. If this is your first time, you will need to configure the worker nodes on your cluster.

Terminal

Status

Cluster name

Disk status:

Worker status

Service status

External Log

Cluster status log

Initial Cluster Configuration

Welcome to Galaxy 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 specify the size of your persistent data storage, if any.

Start a full Galaxy Cluster. Specify initial storage size (in Gigabytes)

GB

[Show more startup options](#)

Start Cluster



Master node is online



Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application will allow you to manage this cloud instance and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be able to add and remove additional services as well as 'worker' nodes on which jobs are run.

[Terminate cluster](#)[Add nodes ▼](#)[Remove nodes](#)[Access Galaxy](#)

Status

Cluster name: local test**Disk status:** 0 / 0 (0%)**Worker status:** Idle: 0 Available: 0 Requested: 0**Service status:** Applications 🟡 Data @**External Logs:**[Cluster status log](#)**Autoscaling is off.**
Turn on?



Add extra worker nodes

Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application will allow you to manage this cloud instance and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be able to add and remove additional services as well as 'worker' nodes on which jobs are run.

Terminate cluster Add nodes ▼ Remove nodes Access Galaxy

Status

Cluster name: local test
Disk status: 0 / 0 (0%)
Worker status: Idle: 0 Av
Service status: Application
External Logs:

Cluster status log

Add nodes

Number of nodes to start:

Type of Nodes(s):

Start Additional Nodes

Autoscaling is off.
Turn on?



New instances are pending



HPC Cloud Management Console

vm overview

vm configuration

disk image upload

disk image management

hosts

networks

public firewall exceptions

quotas

Logged in as mdhollander - logout | version: 1.0.1



last refresh was 3 seconds ago: [refresh now]

Deploy a new VM

Cloud vm's:

Id	User	Name	VM State	LCM State	Memory	Host	VNC Port	Time	Links	Selection	resume ok
4579	mdhollander	Galaxy_Main	stopped	init	1,024 MB	node15-one	10479	107d 21:20:45	[console] [details] [log]	<input type="checkbox"/>	
4858	mdhollander	Galaxy_master	active	running	1,024 MB	node14-one	10758	70d 03:14:24	[console] [details] [log]	<input type="checkbox"/>	
5115	mdhollander	Cloudman_Node	pending	init	0 MB	n/a	11015	0d 00:00:15	[details] [log]	<input type="checkbox"/>	
5116	mdhollander	Cloudman_Node	pending	init	0 MB	n/a	11016	0d 00:00:12	[details] [log]	<input type="checkbox"/>	
5117	mdhollander	Cloudman_Node	pending	init	0 MB	n/a	11017	0d 00:00:09	[details] [log]	<input type="checkbox"/>	



New instances are pending #2



Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application will allow you to manage this cloud instance and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be able to add and remove additional services as well as 'worker' nodes on which jobs are run.

[Terminate cluster](#)
[Add nodes ▼](#)
[Remove nodes ▼](#)
[Access Galaxy](#)

Status

Cluster name: local test

Disk status: 0 / 0 (0%)

Worker status: Idle: 0 Available: 0 Requested: 3

Service status: Applications ● Data @

External Logs: [Galaxy Log](#)



Autoscaling is **off**.
Turn on?

Cluster status log



New instances are running



HPC Cloud Management Console

vm overview

vm configuration

disk image upload

disk image management

hosts

networks

public firewall exceptions

quotas

Logged in as mdhollander - logout | version: 1.0.1



last refresh was 2 seconds ago: [refresh now]



Deploy a new VM

Cloud vm's:

Id	User	Name	VM State	LCM State	Memory	Host	VNC Port	Time	Links	Selection
4579	mdhollander	Galaxy_Main	stopped	init	1,024 MB	node15-one	10479	107d 21:22:22	[console] [details] [log]	<input type="checkbox"/>
4858	mdhollander	Galaxy_master	active	running	1,024 MB	node14-one	10758	70d 03:16:00	[console] [details] [log]	<input type="checkbox"/>
5115	mdhollander	Cloudman_Node	active	running	1,024 MB	node11-one	11015	0d 00:01:51	[console] [details] [log]	<input type="checkbox"/>
5116	mdhollander	Cloudman_Node	active	running	1,024 MB	node14-one	11016	0d 00:01:48	[console] [details] [log]	<input type="checkbox"/>
5117	mdhollander	Cloudman_Node	active	running	1,024 MB	node16-one	11017	0d 00:01:45	[console] [details] [log]	<input type="checkbox"/>

resume ▾

ok



New instances are online



Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application allows you to manage this instance of Galaxy CloudMan. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to add and remove 'worker' nodes for running jobs.

[Terminate cluster](#)[Add nodes ▼](#)[Remove nodes ▼](#)[Access Galaxy](#)

Status

Cluster name: local test

Disk status: 0 / 0 (0%)

Worker status: Idle: 0 Available: 4 Requested: 4

Service status: Applications ● Data ●

External Logs: [Galaxy Log](#)



Autoscaling is **off**.
Turn on?

[Cluster status log](#)



Galaxy is accessible



```
galaxy@ubuntu:/home/cloud$ qstat -f
queue name          qtype resv/used/tot. load_avg arch      states
-----
all.q@ubuntu        BIP   0/0/1           0.26  lx24-and64
all.q@worker-5118   BIP   0/0/1           0.05  lx24-and64
all.q@worker-5119   BIP   0/0/1           0.00  lx24-and64
all.q@worker-5120   BIP   0/0/1           0.08  lx24-and64
all.q@worker-5121   BIP   0/0/1           0.04  lx24-and64
```

```
SNP/WGA: galaxy@ubuntu:/home/cloud$ qstat
Human_Gen job-ID prior name user state submit/start at queue slots ja-task-ID
VCF Tools
EMBOSS 1 0.55500 galaxy 117 galaxy r 09/01/2011 15:49:02 all.q@ubuntu 4
```

```
## Compute
sequence length on d 29
## Compute
sequence length on d 29
```

