

Developing distributed analysis pipelines with shared community resources using CloudBioLinux and CloudMan

Brad Chapman
Bioinformatics Core
Harvard School of Public Health

Acknowledgements

CloudBioLinux – Ntino Krampis, Tim
Booth, Dawn Field, Pjotr Prins and
CloudBioLinux community

CloudMan – Enis Afgan, James Taylor

Exome pipeline – HSPH, MGH, Win Hide,
Oliver Hofmann

Follow along

<http://www.slideshare.net/chapmanb>

Cue the “lots of data” slide

```
ls -lh fastq/
24G  1_110907_AD08A5ACXX_1_fastq.txt
21G  1_110907_AD08A5ACXX_2_fastq.txt
24G  2_110907_AD08A5ACXX_1_fastq.txt
20G  2_110907_AD08A5ACXX_2_fastq.txt
```

Rapidly changing tools

Whole genome, deep coverage v1

Warning: the material on this page is considered out of date by the GSA team.

Best Practice Variant Detection with the GATK v2

Warning: the material on this page is considered out of date by the GSA team.

Best Practice Variant Detection with the GATK v3

Science – fundamental challenge

75% one-off experimental
25% reused code

Unfortunate result

▲ Why scientific programming does not compute (nature.com)

188 points by szany 67 days ago | comments

▲ ajdecon 67 days ago | link

(Disclaimer: my background is in materials physics, and it may be different in other fields. But I doubt it.)

Unfortunately there is very little *direct* incentive for research scientists to write or publish clean, readable code:

<http://news.ycombinator.com/item?id=2735537>

Hard choices

Computation

Demands flexible, well-architected, scalable code

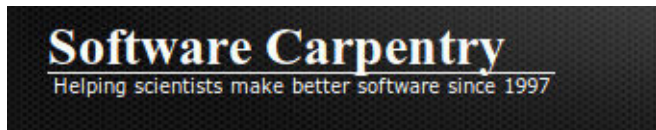
Science

Requires rapid turn around and experimentation

2 solutions (at least)

- 1 Improve your programming skills
- 2 Utilize community resources

Become a better coder



<http://software-carpentry.org/>

Community resources

- Share painful parts
- Base of well-written, scalable code

Start each problem from a higher level of abstraction

Community components

- CloudBioLinux – install software
- CloudMan – manage cluster
- Exome analysis pipeline – do science

CloudBioLinux

- Amazon image with bioinformatics software and libraries
- Automated build framework
- Community effort to maintain and extend

<http://cloudbiolinux.org>

CloudMan

- SGE cluster plus automation
- Web interface and monitoring
- Persistence and sharing
- Powers the Galaxy Cloud offering

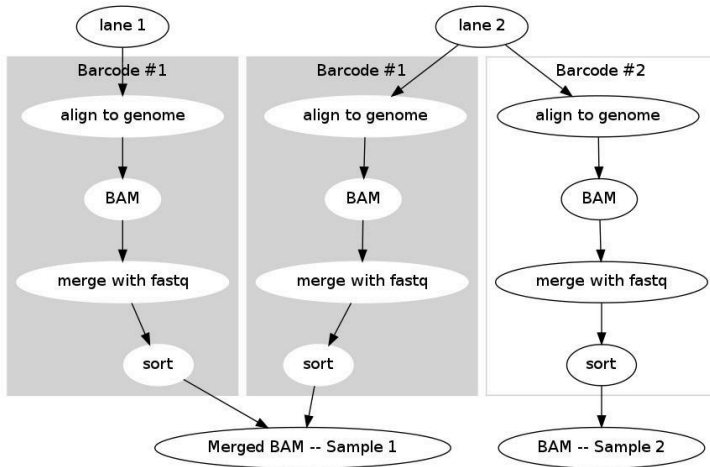
<http://wiki.g2.bx.psu.edu/Admin/Cloud>

Exome analysis pipeline

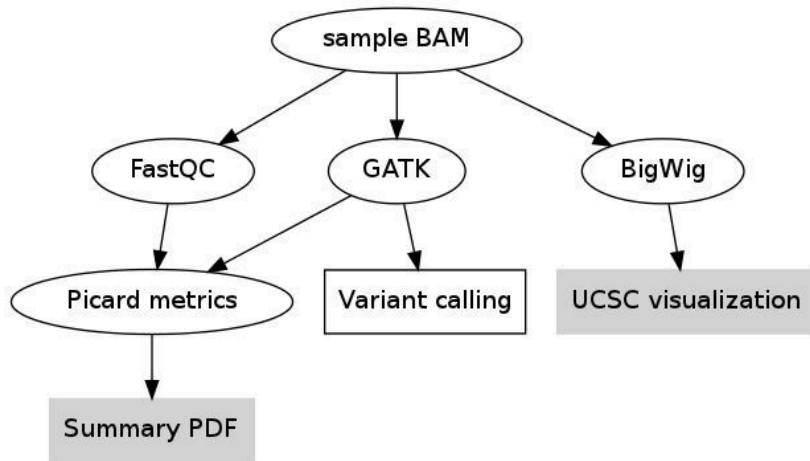
- Existing algorithms
 - Aligners – Bowtie, BWA
 - Variation – GATK
 - Quality assessment – FastQC, Picard
- Messaging system – AMQP

[https://github.com/chapmanb/bcbb/
tree/master/nextgen](https://github.com/chapmanb/bcbb/tree/master/nextgen)

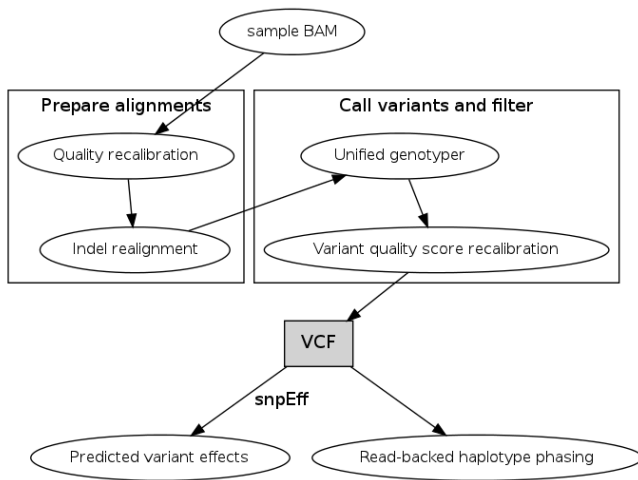
Fastq lane processing



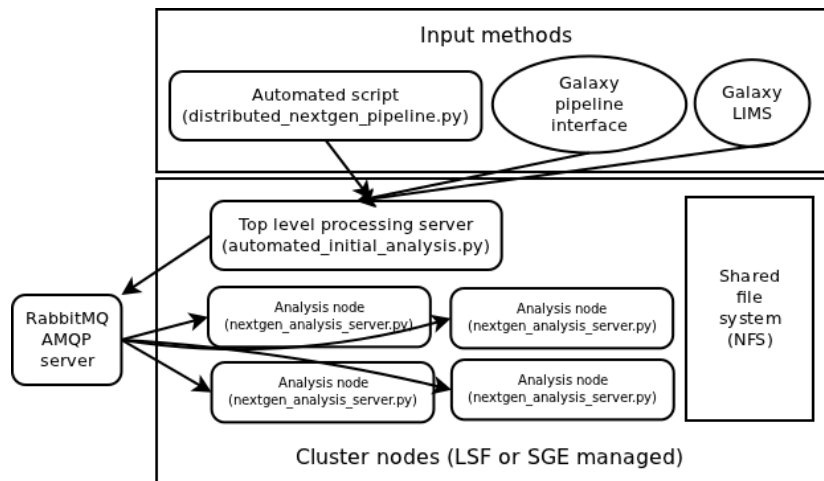
Sample processing

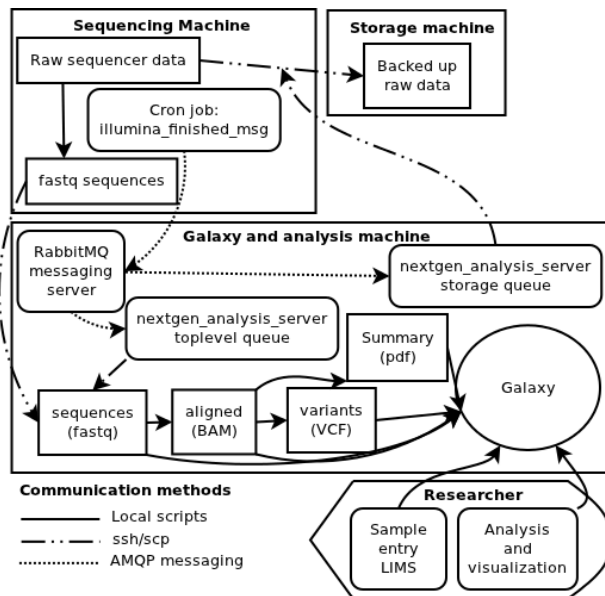


Variant calling



Parallelization





Amazon

- Virtual machines
 - Share
 - Reproduce
 - Coordinate
- Accessibility



What are we going to do?

- Use AWS console to boot CloudBioLinux
- Setup CloudMan in AWS console
- Boot CloudMan instance with demo data

What are we going to do? continued

- Manage cluster with CloudMan interface
- Setup messaging queue
- Run pipeline, examine results
- Share cluster

CloudBioLinux

- Select and launch CloudBioLinux AMI from AWS console
- Connect
 - FreeNX graphical client
 - ssh

Full tutorial PDF: <http://j.mp/nnh5TE>

Prep work

- Signup for AWS account:

<http://aws.amazon.com/>

- Create login key pair in AWS Console
- Install NX client:

<http://www.nomachine.com/select-package-client.php>



Navigation
Region:
US East (Virginia)
EC2 Dashboard
INSTANCES
Instances
Spot Requests
Reserved Instances
IMAGES
AMIs
Bundle Tasks
ELASTIC BLOCK STORE
Volumes
Snapshots
NETWORK & SECURITY
Security Groups
Elastic IPs
Placement Groups
Load Balancers
Key Pairs

Amazon EC2 Console Dashboard
Getting Started
To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.
[Launch Instance](#)
Note: Your instances will launch in the US East (Virginia) region.
Service Health
Service Status

Current Status	Details
✓ Amazon EC2 (US East - N. Virginia)	Service is operating normally View complete service health details

Availability Zone Status

Current Status	Details
✓ us-east-1a	Availability zone is operating normally

My Resources
You are using the following Amazon EC2 resources in the US East (Virginia) region: [Refresh](#)
0 Running Instances 0 Elastic IPs
1 EBS Volume 3 EBS Snapshots
1 Key Pair 2 Security Groups
0 Load Balancers 0 Placement Groups
Related Links
[Documentation](#)
[All EC2 Resources](#)
[Forums](#)
[Feedback](#)
[Report an Issue](#)

<https://console.aws.amazon.com/ec2/>

AWS Management Console | **Amazon Elastic Compute Cloud (EC2)** | Brad Chapman | Help

Request Instances Wizard

Cancel X

CHOOSE AN AMI | INSTANCE DETAILS | CREATE KEY PAIR | CONFIGURE FIREWALL | REVIEW

Choose an Amazon Machine Image (AMI) from one of the tabbed lists below by clicking its **Select** button.

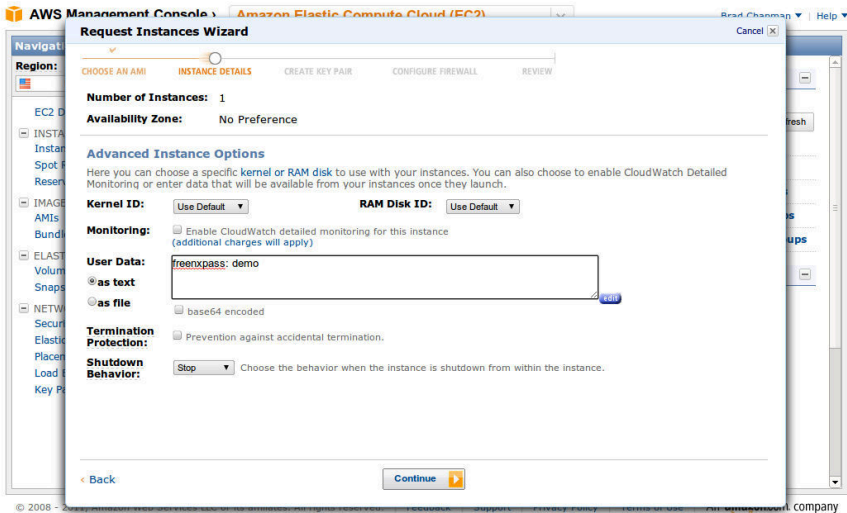
Quick Start | **My AMIs** | **Community AMIs**

Viewing: All Images CloudBioLinux 1 to 2 of 2 Items

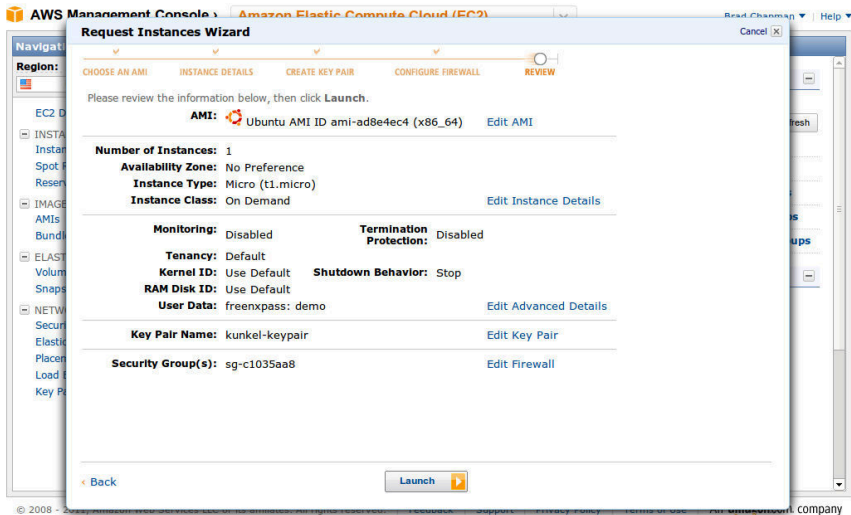
AMI ID	Root Device	Manifest	Platform	
ami-90d32af9	ebs	678711657553/CloudBioLinux Ubuntu 10.04 LTS 64	Ubuntu	Select
ami-ad8e4ec4	ebs	678711657553/CloudBioLinux Ubuntu 11.04 64bit 20	Ubuntu	Select

★ Free tier eligible if used with a micro instance. See [AWS free tier](#) for complete details and terms.

Select CloudBioLinux image from Community AMIs



enter NX password in user-data (freenxpass: secret)



Launch CloudBioLinux server

AWS Management Console > **Amazon Elastic Compute Cloud (EC2)** Brad Chapman | Help

Navigation

Region:
US East (Virginia)

- EC2 Dashboard
- INSTANCES
 - Instances**
 - Spot Requests
 - Reserved Instances
- IMAGES
 - AMIs
 - Bundle Tasks
- ELASTIC BLOCK STORE
 - Volumes
 - Snapshots
- NETWORK & SECURITY
 - Security Groups
 - Elastic IPs
 - Placement Groups
 - Load Balancers
 - Key Pairs

My Instances

[Launch Instance](#) [Instance Actions](#) [Show/Hide](#) [Refresh](#) [Help](#)

Viewing: All Instances All Instance Types Search 1 to 1 of 1 Instances

	Name	Instance	AMI ID	Root Device	Type	Status	Security Groups	Key Pair Name
☒	empty	i-24903744	ami-ad8e4ec4	ebs	t1.micro	running	CloudBioLinux	kunkel-keypair

Block Devices: [sda1](#)

Public DNS: ec2-184-73-64-47.compute-1.amazonaws.com

Private DNS: ip-10-244-166-173.ec2.internal

Private IP Address: 10.244.166.173

Launch Time: 2011-09-13 09:23 EDT

State Transition Reason:

Termination Protection: Disabled

Get external hostname from Instances page

Session

NOMACHINE

Insert name of the session. Your configuration settings will be saved with this name.

Session

Insert server's name and port where you want to connect.

Host Port

Select type of your internet connection.

MODEM ISDN ADSL WAN LAN

< Back Next > Cancel

Desktop

NOMACHINE

Using NX Client you can run RDP, VNC and X desktops, depending on what the service provider has made available.

Unix GNOME Settings...

Select size of your remote desktop.

1024x768 W: 800 H: 600

Authorization credentials are always encrypted at the time connection is established. To enhance performance, you can disable the encryption of the data traffic.

☐ Disable encryption of all traffic

< Back Next > Cancel

NOMACHINE

Login

Password

Session

☐ Login as a guest user

Configure... Login Close

Connect using NX client, with ubuntu user and secret password

- Accessories >
- Bioinformatics >
- Internet >
- Programming >
- Science >
- Sound & Video >
- System Tools >
- Ubuntu Software Center

```
ubuntu@ip-10-244-166-173: ~
File Edit View Search Terminal Help

Program: bwa (alignment via Burrows-Wheeler transformation)
Version: 0.5.9-r16
Contact: Heng Li <lh3@sanger.ac.uk>

Usage:  bwa <command> [options]

Command: index          index sequences in the FASTA format
aln                  gapped/ungapped alignment
samse               generate alignment (single ended)
sampe               generate alignment (paired ended)
bwsw                BWA-SW for long queries

fa2pac              convert FASTA to PAC format
pac2bwt             generate BWT from PAC
pac2bwtgen          alternative algorithm for generating BWT
bwtupdate            update .bwt to the new format
pac_rev             generate reverse PAC
bwt2sa              generate SA from BWT and Occ
pac2cspac            convert PAC to color-space PAC
stdsw               standard SW/NW alignment

ubuntu@ip-10-244-166-173:~$
```



J. Craig VenterTM
I N S T I T U T E



**NATURAL
ENVIRONMENT
RESEARCH COUNCIL**


```
kunkel:~ $ ssh -i ~/.ec2/id-kunkel.keypair ubuntu@ec2-184-73-64-47.compute-1.amazonaws.com
Warning: Permanently added 'ec2-184-73-64-47.compute-1.amazonaws.com' (RSA) to the list of known hosts.
Welcome to Ubuntu 11.04 (GNU/Linux 2.6.38-8-virtual x86_64)

 * Documentation:  https://help.ubuntu.com/

System information disabled due to load higher than 1.0
-----
At the moment, only the core of the system is installed. To tune the
system to your needs, you can choose to install one or more
predefined collections of software by running the following
command:

    sudo tasksel --section server
-----

47 packages can be updated.
26 updates are security updates.

Last login: Tue Sep 13 13:40:34 2011 from 209-6-39-30.c3-0.smr-ubr1.sbo-smr.ma.cable.rcn.com
ubuntu@ip-10-244-166-173:~$ bowtie
No index, query, or output file specified!
Usage:
  bowtie [options]* <ebwt> {-1 <m1> -2 <m2> | --12 <r> | <s>} [<hit>]
```

Connect with ssh, using private ssh key-pair

AWS Management Console | **Amazon Elastic Compute Cloud (EC2)** | Brad Chapman | Help

Navigation

Region: US East (Virginia)

- EC2 Dashboard
- INSTANCES
 - Instances**
 - Spot Requests
 - Reserved Instances
- IMAGES
 - AMIs
 - Bundle Tasks
- ELASTIC BLOCK STORE
 - Volumes
 - Snapshots
- NETWORK & SECURITY
 - Security Groups
 - Elastic IPs
 - Placement Groups
 - Load Balancers
 - Key Pairs

My Instances

Launch Instance Instance Actions Show/Hide Refresh Help

Viewing: All Instances All Instance Types Search 1 to 1 of 1 Instances

Name	Instance	AMI ID	Root Device	Type	Status	Security Groups	Key Pair Name
✓ empty	I-24903744	ami-ad8e4ec4	ebs	t1.micro	running	CloudBioLinux	kunkel-keypair

Instance Management

- Connect
- Get System Log
- Create Image (EBS AMI)
- Add/Edit Tags
- Change Security Groups
- Change Source / Dest Check
- Launch More Like This
- Disassociate IP Address
- Change Termination Protection
- View/Change User Data
- Change Instance Type
- Change Shutdown Behavior

Instance Lifecycle

- Terminate**
- Reboot
- Stop
- Start

CloudWatch Monitoring

- Enable Detailed Monitoring
- Disable Detailed Monitoring

Block Devices: sda1

Public DNS: ec2-184-73-64-47.compu

Private DNS: ip-10-244-166-173.ec2.ir

Private IP Address: 10.244.166.173

Launch Time: 2011-09-13 09:23 EDT

State Transition Reason:

Termination Protection: Disabled

Terminate the server when finished

Setup CloudMan in AWS console

- Create a custom security group

Full tutorial:

<http://wiki.g2.bx.psu.edu/Admin/Cloud>

AWS Management Console | **Amazon Elastic Compute Cloud (EC2)** | Brad Chapman | Help

Navigation

Region: US East (Virginia)

- EC2 Dashboard
- INSTANCES
 - Instances
 - Spot Requests
 - Reserved Instances
- IMAGES
 - AMIs
 - Bundle Tasks
- ELASTIC BLOCK STORE
 - Volumes
 - Snapshots
- NETWORK & SECURITY
 - Security Groups**
 - Elastic IPs
 - Placement Groups
 - Load Balancers
 - Key Pairs

Security Groups

Create Security Group Delete Show/Hide Refresh Help

Viewing: EC2 Security Groups Search 1 to 2 of 2 Items

	Name	VPC ID	Description
☐			
☒			

Create Security Group Cancel

Name: CloudBioLinux

Description: CloudBioLinux

VPC: No VPC

Cancel Yes, Create

Security Group selected

Security Group: default

Details Inbound

Group Name: default

Group ID: sg-4752b62e

Group Description: default group

VPC ID: -

Create security group rules following wiki instructions

AWS Management Console > **Amazon Elastic Compute Cloud (EC2)** Brad Chapman ▾ Help ▾

Navigation
Region:
US East (Virginia) ▾

[EC2 Dashboard](#)

INSTANCES

[Instances](#)
[Spot Requests](#)
[Reserved Instances](#)

IMAGES

[AMIs](#)
[Bundle Tasks](#)

ELASTIC BLOCK STORE

[Volumes](#)
[Snapshots](#)

NETWORK & SECURITY

[Security Groups](#)
[Elastic IPs](#)
[Placement Groups](#)
[Load Balancers](#)
[Key Pairs](#)

Security Groups

Create Security Group
Delete
Show/Hide
Refresh
Help

Viewing: EC2 Security Groups ▾

1 to 2 of 2 Items

	Name	VPC ID	Description
☒	CloudBioLinux		CloudBioLinux
☐	default		default group

1 Security Group selected
Security Group: CloudBioLinux

Details
Inbound

Create a new rule:
Custom TCP rule ▾
Port range:

(e.g., 80 or 49152-65535)
Source:

(e.g., 192.168.2.0/24, sg-47ad482e, or 1234567890/default)
Add Rule

TCP Port (Service)	Source	Action
0 - 65535	sg-c1035aa8 (CloudBioLinux)	Delete
20 - 21	0.0.0.0/0	Delete
22 (SSH)	0.0.0.0/0	Delete
80 (HTTP)	0.0.0.0/0	Delete
30000 - 30100	0.0.0.0/0	Delete
42284	0.0.0.0/0	Delete

Final security group specifications

Navigation icons: back, forward, search, etc.

Boot CloudMan instance with demo data

- Start server
- Pass in CloudMan user data
- Load shared CloudMan image

AWS Management Console | **Amazon Elastic Compute Cloud (EC2)** | Brad Chapman | Help

Request Instances Wizard

Cancel

Region:

Navigation: EC2 Dashboard | INSTANCES | Spot Instances | Reserved Instances | IMAGES | AMIs | Bundles | ELASTIC VOLUMES | Snapshots | NETWORKING | Security Groups | Elastic Network Interfaces | Placement Groups | Load Balancing | Key Pairs

Steps: CHOOSE AN AMI | INSTANCE DETAILS | CREATE KEY PAIR | CONFIGURE FIREWALL | REVIEW

Choose an Amazon Machine Image (AMI) from one of the tabbed lists below by clicking its **Select** button.

Quick Start | **My AMIs** | **Community AMIs**

Viewing: All Images | CloudBioLinux | 1 to 2 of 2 Items

AMI ID	Root Device	Manifest	Platform	
ami-90d32af9	ebs	678711657553/CloudBioLinux Ubuntu 10.04 LTS 64bit	Ubuntu	Select
ami-ad8e4ec4	ebs	678711657553/CloudBioLinux Ubuntu 11.04 64bit 2D	Ubuntu	Select

★ Free tier eligible if used with a micro instance. See [AWS free tier](#) for complete details and terms.

Follow same procedure as CloudBioLinux

Create CloudMan user-data file

```
cluster_name: cbldemo  
password: cbl  
access_key: your_access_key  
secret_key: your_long_AWS_secret_key
```


AWS Management Console Amazon Elastic Compute Cloud (EC2) Brad Chapman Help

Request Instances Wizard

Cancel X

CHOOSE AN AMI INSTANCE DETAILS CREATE KEY PAIR CONFIGURE FIREWALL REVIEW

Number of Instances: 1

Availability Zone: us-east-1c

Advanced Instance Options

Here you can choose a specific [kernel](#) or [RAM disk](#) to use with your instances. You can also choose to enable CloudWatch Detailed Monitoring or enter data that will be available from your instances once they launch.

Kernel ID: Use Default ▼ **RAM Disk ID:** Use Default ▼

Monitoring: ☐ Enable CloudWatch detailed monitoring for this instance (additional charges will apply)

User Data: Choose File cloudman-cbldemo.txt

☒ as text

☒ as file ☐ base64 encoded

Termination Protection: ☐ Prevention against accidental termination.

Shutdown Behavior: Stop ▼ Choose the behavior when the instance is shutdown from within the instance.

Back Continue

Provide user-data from file

AWS Management Console Amazon Elastic Compute Cloud (EC2) Brad Chapman Help

Request Instances Wizard

Cancel

Region:  **US**

EC2 D

- INSTA
- Instan
- Spot F
- Reser
- IMAGE
- AMIs
- Bundl
- ELAST
- Volum
- Snaps
- NETW
- Secur
- Elastic
- Placen
- Load B
- Key Pa

Navigation: CHOOSE AN AMI | INSTANCE DETAILS | CREATE KEY PAIR | **CONFIGURE FIREWALL** | REVIEW

Security groups determine whether a network port is open or blocked on your instances. You may use an existing security group, or we can help you create a new security group to allow access to your instances using the suggested ports below. Add additional ports now or update your security group anytime using the Security Groups page.

☒ **Choose one or more of your existing Security Groups**

sg-c1035aa8 - CloudBioLinux

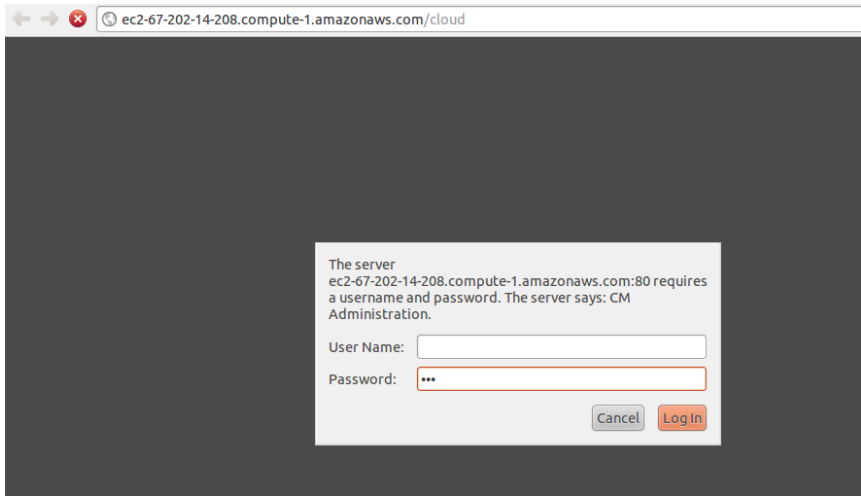
sg-4752b62e - default

(Selected groups: sg-c1035aa8)

☐ Create a new Security Group

[Back](#) [Continue](#)

Choose created security group



Login to instance with password from user-data

CloudMan share-an-instance

- Persist data in a CloudMan cluster
- Easily sharable

For this demo

cm-b53c6f1223f966914df347687f6fc818/shared/2011-10-07-14-00

ec2-67-202-14-208.compute-1.amazonaws.com/cloud#

Galaxy Cloudman

Admin | Report bugs | Wiki | Screenshot

Galaxy Cloudman

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.

Initial Cluster Configuration

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

GB

☒ Share-an-instance

Shared instance bucket path

☐ Data volume and SGE only. Specify initial storage size (in Gigabytes)

GB

☐ SGE Only. No persistent storage created.

[Hide extra options](#)

Start Cluster

Import shared instance with demo data

Manage cluster with CloudMan

- Web-based console
- Monitor running processes
- Add nodes to cluster as needed

Galaxy Cloudman Console

Warning: You are running out of disk space. Use the disk icon below to increase your volume size.

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: cbldemo 

Disk status: 17G / 20G (83%) 

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

Service status: Applications  Data 



Autoscaling is **off**.
Turn on?

Cluster status log

```
17:36:40 - Master starting
17:36:42 - Completed initial cluster configuration.
17:37:02 - Prerequisites OK; starting service 'SGE'
17:37:08 - Configuring SGE...
17:37:16 - Successfully setup SGE; configuring SGE
17:37:17 - Saved file 'persistent_data.yaml' to bucket 'cm-56133fba0111c9ec3841ed40368b64e6'
17:37:17 - Saved file 'cm_boot.py' to bucket 'cm-56133fba0111c9ec3841ed40368b64e6'
17:37:17 - Saved file 'cm.tar.gz' to bucket 'cm-56133fba0111c9ec3841ed40368b64e6'
17:37:17 - Saved file 'cbldemo.clusterName' to bucket 'cm-56133fba0111c9ec3841ed40368b64e6'
17:45:09 - Retrieved file 'shared/2011-08-19--21-00/shared_instance_file_list.txt' from bucket 'cm-0011923649e9271f17c4f83ba6846db0' to 'shared_instance_file_list.txt'.
17:45:09 - Retrieved file 'persistent_data.yaml' from bucket 'cm-56133fba0111c9ec3841ed40368b64e6' to
```

CloudMan console to interact with cluster

Galaxy Cloudman Console

Warning: You are running out of disk space. Use the disk icon below to increase your volume size.

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: cbldemo

Disk status: 17G / 20G

Worker status: Idle: 0 Av

Service status: Application

Add nodes

Number of nodes to start:

OK

Type of node(s):

(master node type: m1.large)

Same as Master ▼

Start Additional Nodes

Autoscaling is off.
Turn on?

Cluster status log

17:36:40 - Master starting

17:36:42 - Completed initial cluster configuration.

Add node to cluster

Setup messaging communication

- Command line access to server
- Adjust RabbitMQ configuration
- Setup messaging queue

Command line access to server

```
ssh -i ~/.ec2/id-kunkel.keypair  
ubuntu@ec2-67-202-14-208.compute-1.amazonaws.com
```

Follow approach used to connect to
CloudBioLinux cluster; can also connect via
NX

Edit /export/data/galaxy/universe_wsgi.ini
configuration file to add internal host name.

```
[galaxy_amqp]
host = ip-10-125-10-182.ec2.internal
port = 5672
userid = biouser
password = tester
```

Setup messaging queue

```
$ sudo rabbitmqctl add_user biouser tester
creating user 'biouser' ...
...done.
$ sudo rabbitmqctl add_vhost bionextgen
creating vhost 'bionextgen' ...
...done.
$ sudo rabbitmqctl set_permissions -p bionextgen
    biouser ".*" ".*" ".*"
setting permissions for user 'biouser' in vhost 'bionextgen' ..
...done.
```

Run pipeline, examine results

- Ready to run distributed pipeline
- Demo data – two paired end fastq lanes
- Variant calling workflow

Input sequence data

```
$ ls -1 /export/data/exome_example/fastq/  
7_100326_FC6107FAAXX_1-chr22.fastq  
7_100326_FC6107FAAXX_2-chr22.fastq  
8_100326_FC6107FAAXX_1-chr22.fastq  
8_100326_FC6107FAAXX_2-chr22.fastq
```

Run level: YAML Configuration

```
$ cat /export/data/exome_example/config/run_info.yaml
---
fc_date: '100326'
fc_name: FC6107FAAXX
details:
  - files: [7_100326_FC6107FAAXX_1-chr22.fastq,
            7_100326_FC6107FAAXX_2-chr22.fastq]
    lane: 7
    description: Test replicate 1
    analysis: SNP calling
    genome_build: hg19
    algorithm:
      quality_format: Standard
      hybrid_bait: hybrid_selection/baits.bed
      hybrid_target: hybrid_selection/targets.bed
```

System level: YAML Configuration

```
$ cat /export/data/galaxy/post_process.yaml
```

```
---
```

```
program:
```

```
  bowtie: bowtie
```

```
  bwa: bwa
```

```
  ucsc_bigwig: wigToBigWig
```

```
  picard: /usr/share/java/picard
```

```
  gatk: /usr/share/java/gatk
```

```
  snpEff: /usr/share/java/snpeff
```

```
  fastqc: fastqc
```

```
distributed:
```

```
  cluster_platform: sge
```

```
  platform_args: '-q all.q'
```

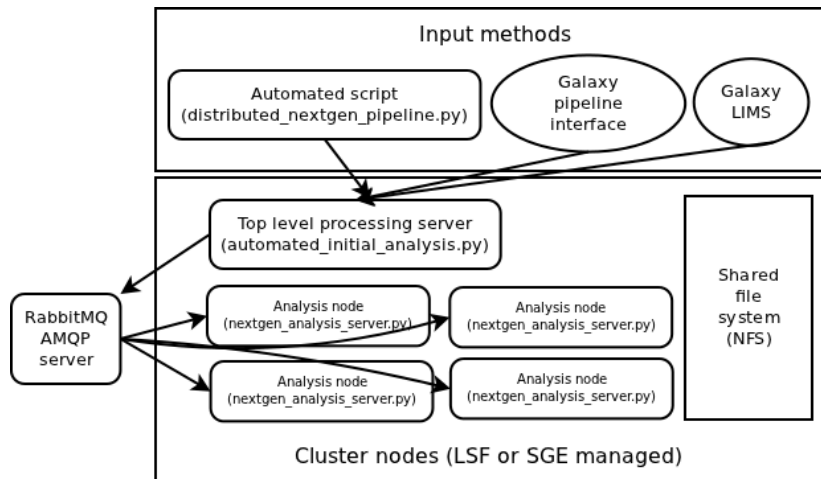
```
  cores_per_host: 1
```

```
  rabbitmq_vhost: bionextgen
```


Run exome pipeline

```
$ cd /export/data/work
$ distributed_nextgen_pipeline.py
  /export/data/galaxy/post_process.yaml
  /export/data/exome_example/fastq
  /export/data/exome_example/config/run_info.yaml
```

What just happened?



Monitoring: SGE queues

```
$ qstat
```

```
ob-ID  prior name state submit/start at    queue
```

```
-----  
1 0.55500 nextgen_an r 18:16:32  all.q@ip-10-125-10-182.ec2.int  
2 0.55500 nextgen_an r 18:16:32  all.q@ip-10-86-254-105.ec2.int  
3 0.55500 automated_ r 18:16:47  all.q@ip-10-125-10-182.ec2.int
```

Monitoring: Analysis directory

```
$ cd /export/data/work
$ ls -lh
drwxr-xr-x  4.0  alignments
-rw-r--r-- 2.0K  automated_initial_analysis.py.o11
drwxr-xr-x   33  log
-rw-r--r-- 15K  nextgen_analysis_server.py.o10
-rw-r--r-- 15K  nextgen_analysis_server.py.o9
drwxr-xr-x 102  tmp
```

Monitoring: Log files

```
$ less nextgen_analysis_server.py.o10
INFO: nextgen_pipeline: Processing sample: Test replicate 2;
    lane 8; reference genome hg19; researcher ;
    analysis method SNP calling
INFO: nextgen_pipeline:
    Aligning lane 8_100326_FC6107FAAXX with bwa aligner
INFO: nextgen_pipeline:
    Combining and preparing wig file [u'', u'Test replicate 2']
INFO: nextgen_pipeline:
    Recalibrating [u'', u'Test replicate 2'] with GATK
```

Retrieve results: Copy files

```
$ upload_to_galaxy.py  
  /export/data/galaxy/post_process.yaml  
  /export/data/exome_example/fastq  
  /export/data/work  
  /export/data/exome_example/config/run_info.yaml
```

Final files copied into new directory; allows cleanup of analysis directory

Retrieve results: Output directory

```
$ ls -lh /export/data/galaxy/storage/100326_FC6107FAAXX/7
-rw-r--r-- 38M 7_100326_FC6107FAAXX.bam
-rw-r--r-- 22M 7_100326_FC6107FAAXX-coverage.bigwig
-rw-r--r-- 72M 7_100326_FC6107FAAXX-gatkrecal.bam
-rw-r--r-- 109K 7_100326_FC6107FAAXX-snp-effects.tsv
-rw-r--r-- 827K 7_100326_FC6107FAAXX-snp-filter.vcf
-rw-r--r-- 1.6M 7_100326_FC6107FAAXX-summary.pdf
```

Share results

- Share-an-instance
- Uses CloudMan web interface
- Reproducible research
 - CloudBioLinux AMI – software
 - CloudMan – data and configuration

Galaxy Cloudman Console

Warning: You are running out of disk space. Use the disk icon

Welcome to Galaxy Cloudman. This application will allow you to manage within. If this is your first time running this cluster, you will need to set up a storage is configured, default services will start and you will be able to add 'worker' nodes on which jobs are run.

Terminate cluster

Add nodes ▼

Remove

Status

Cluster name: cbldemo



Share this cluster instance

Disk status: 17G / 20G (83%) 

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

Service status: Applications  Data 

CloudMan console enables push button sharing



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

Share-an-instance

Can make public or available to specific collaborators

 Galaxy Cloudman

Admin | [Report bugs](#) | [Wiki](#) | [Support](#)

EC2 Cluster Configuration

Are you sure you want to power the cluster off?

This action will shut down all services on the cluster and terminate any worker nodes (instances) associated with this cluster. By default, the master instance will be left alive and should be terminated manually (using the AWS console).

Automatically terminate the master instance?

☒ If checked, this master instance will automatically terminate after all services have been shut down.

Also delete this cluster?

☐ If checked, this cluster will be deleted. **This action is irreversible!** All your data will be deleted.

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

Service status: Applications ● Data ●

Turn on?

When finished, turn everything off through CloudMan

Summary

CloudBioLinux

- Shared machine image of biological software
- Boot from AWS console
- Connect with NX graphical client and ssh

Summary

CloudMan

- Cluster setup and management
- Boot from share-an-instance
- Manage cluster through web interface
- Share final results

Summary

Exome pipeline

- Parallel framework for running analyses
- Run using automated scripts
- Extract alignments, variant calls and summary information

Future: interfaces make it easier

The screenshot shows a web-based workflow interface. On the left is a sidebar with a menu: 'Analysis type' (highlighted in yellow), 'Datasets', 'Parameters', and 'Summary'. The main area has a header 'Alignment' and a sub-header 'Exome variant calling' (highlighted in orange). A dark grey tooltip box on the right contains the text: 'Align reads with BWA, recalibration, realign and variant call with GATK. Returns sorted BAM file, variant calls in VCF format, tab separated file of predicted effects and PDF with quality statistics.' At the bottom right of the main area is a 'Next' button.

<https://bitbucket.org/hbc/galaxy-central-hbc>

Future: Simplified file selection

Analysis type

Datasets

Parameters

Summary

Available

1_110824_V0P8FM-gatkrecal.bam

To process

Read

7_100326_FC6107FAAXX_1-chr22.fastq

Read pair (optional)

7_100326_FC6107FAAXX_2-chr22.fastq

Previous

Next

Future: Top level parameters

Analysis type

Datasets

Parameters

Summary

Include reads that map to multiple locations?
yes

Target file (from history; bed format)
targets.bed

Bait file (from history; bed format)
baits.bed

Organism
Human Feb. 2009 (GRCh37/hg19) (h...)

Barcodes

Illumina 454-Rapid

1 : ATCAGC
2 : CGATGT
3 : TTAGGC
4 : TGACCA
5 : ACAGTG
6 : GCCAAT
7 : CAGATC
8 : ACTTGA
9 : GATCAG
10 : TAGCTT
11 : GGCTAC
12 : CTCTA

Samples

Add new sub-sample

Previous Next

Future: Galaxy data libraries

Data Library "bchapman@hsph.harvard.edu"

 Name	Message	Uploaded By	Date	File Size
  110824_VOP8FM ▾				
  1 ▾	1			
 1 110824_VOP8FM.bam ▾		bchapman@hsph.harvard.edu	2011-08-25	37.8 Mb
 1 110824_VOP8FM-coverage.bigwig ▾		bchapman@hsph.harvard.edu	2011-08-25	21.9 Mb
 1 110824_VOP8FM-gatkrecal.bam ▾		bchapman@hsph.harvard.edu	2011-08-25	71.6 Mb
 1 110824_VOP8FM-snp-effects.tsv ▾		bchapman@hsph.harvard.edu	2011-08-25	108.8 Kb
 1 110824_VOP8FM-snp-filter.vcf ▾		bchapman@hsph.harvard.edu	2011-08-25	826.2 Kb
 1 110824_VOP8FM-summary.pdf ▾		bchapman@hsph.harvard.edu	2011-08-25	1.6 Mb
For selected datasets: Import to current history Go				

Future: Galaxy analysis

Galaxy Analyze Data Workflow Shared Data Visualization Admin Help User

Tools Options ▾

SCDE TOOLS

[Gene List Comparison](#)

GALAXY 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](#)

[Graph/Display Data](#)

[Multiple regression](#)

[Multivariate Analysis](#)

[FASTA manipulation](#)

: 1 (1.110824.V0P8FM-sort)

Reference organism	hg19	
Total	688,800	76bp paired
Aligned	679,724	(98.7%)
Pairs aligned	678,388	(98.5%)
Alignment combinations	330,194	
Pair duplicates	31,229	(9.2%)
Insert size	151.7	4 / 30.5
On bait bases	21,699,537	(47.3%)
Near bait bases	5,336,058	(11.8%)
Off bait bases	18,249,851	(40.3%)
Mean bait coverage	231.8	
On target bases	17,633,684	(38.9%)
Mean target coverage	244x	
Use coverage targets	93.1%	
Zero coverage targets	4.3%	
Fold enrichment	16056x	
Total variations	68	
In dbSNP	100.0%	
Transition/Transversion (all)	-1.00	
Transition/Transversion (dbSNP)	-1.00	
Transition/Transversion (novel)	-1.00	

Table 1: Summary of lane results

History Options ▾

Unnamed history 218.8 Mb

[17: 1 110824_V0P8FM- summary.pdf](#) 1.6 Mb
format: pdf, database: hg19
Info: uploaded pdf file

[16: 1 110824_V0P8FM- gatkrecal.bam](#)

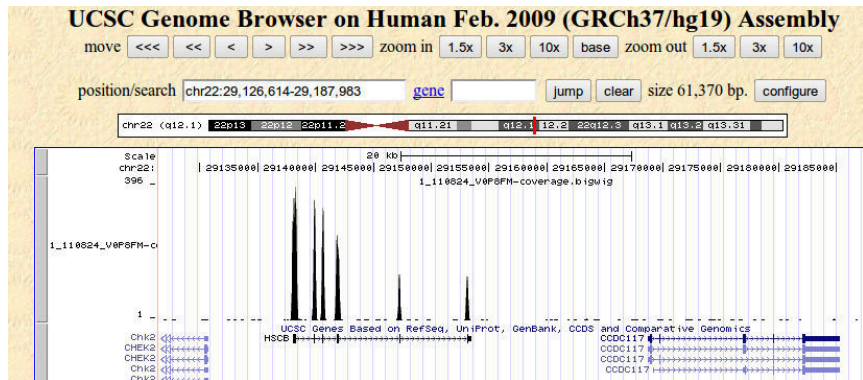
[15: 1 110824_V0P8FM- coverage.bigwig](#) 21.9 Mb
format: bigwig, database: hg19
Info: uploaded bigwig file

display at UCSC [main](#)

Binary UCSC BigWig file

Future: External UCSC

visualization



Read more

- Step-by-step instructions

<http://j.mp/rp69nx>

- Approaches to parallelism

<http://j.mp/nPQHcm>

- Future work

<http://bcbio.wordpress.com>