Bioinformatics tutorials Documentation

Release 1

Jiarong

May 18, 2017

Contents

Contents:

CHAPTER 1

Welcome to 2014 MSU QB summer microbial genomics tutorial

Day	Schedule
Wed 8/13	• 2-3pm: *Get started with Amazon EC2* • 3-5pm: *Read mapping and variant calling*
Thu 8/14	• 2-5pm: *Genome assembly*

Sequencing technologes

- 454
- Illumina
- PacBio
- NanoPore

Resources

- Biostar

  A high quality question & answer Web site.
- SEQanswers

  A discussion and information site for next-generation sequencing.
- Software Carpentry lessons

  A large number of open and reusable tutorials on the shell, programming, version control, etc.

Table Of Contents

Amazon Web Services instructions

Start up an EC2 instance

Here, we're going to startup an Amazon Web Services (AWS) Elastic Cloud Computing (EC2) "instance", or computer.

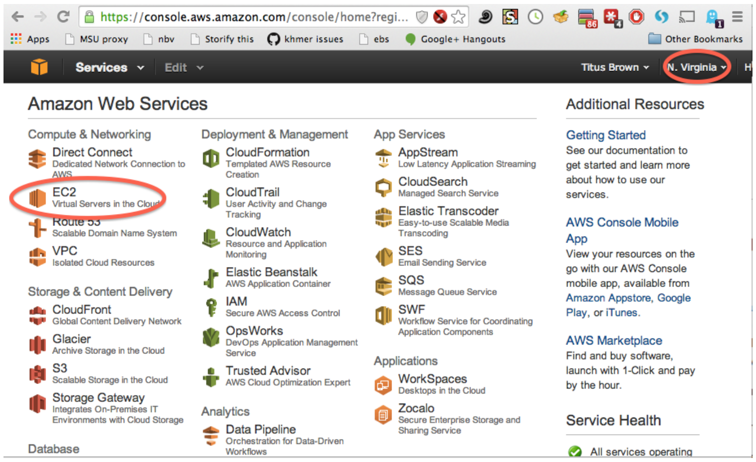
Go to 'https://aws.amazon.com' in a Web browser.

Select 'My Account/Console' menu option 'AWS Management Console."

Log in with the following:

- **username: qb2014msu@gmail.com**
- **password: TemporaryQB2014**

Make sure it says North Virginia in the upper right, then select EC2 (upper left).


https://console.aws.amazon.com/console/home?regi...
Apps
MSU proxy
nbv
Storify this
khmer issues
ebs
Google+ Hangouts
Other Bookmarks
Services
Edit
Titus Brown
N. Virginia
Amazon Web Services
Compute & Networking
Direct Connect
Dedicated Network Connection to AWS
EC2
Virtual Servers in the Cloud
Route 53
Scalable Domain Name System
VPC
Isolated Cloud Resources
Storage & Content Delivery
CloudFront
Global Content Delivery Network
Glacier
Archive Storage in the Cloud
S3
Scalable Storage in the Cloud
Storage Gateway
Integrates On-Premises IT Environments with Cloud Storage
Database
Deployment & Management
CloudFormation
Templated AWS Resource Creation
CloudTrail
User Activity and Change Tracking
CloudWatch
Resource and Application Monitoring
Elastic Beanstalk
AWS Application Container
IAM
Secure AWS Access Control
OpsWorks
DevOps Application Management Service
Trusted Advisor
AWS Cloud Optimization Expert
Analytics
Data Pipeline
Orchestration for Data-Driven Workflows
App Services
AppStream
Low Latency Application Streaming
CloudSearch
Managed Search Service
Elastic Transcoder
Easy-to-use Scalable Media Transcoding
SES
Email Sending Service
SQS
Message Queue Service
SWF
Workflow Service for Coordinating Application Components
Applications
WorkSpaces
Desktops in the Cloud
Zocalo
Secure Enterprise Storage and Sharing Service
Additional Resources
Getting Started
AWS Console Mobile App
AWS Marketplace
Service Health

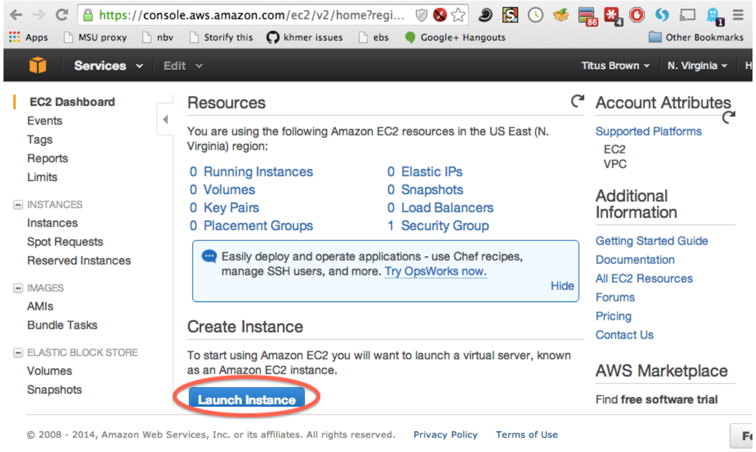
See our documentation to get started and learn more about how to use our services.

View your resources on the go with our AWS Console mobile app, available from Amazon Appstore, Google Play, or iTunes.

Find and buy software, launch with 1-Click and pay by the hour.

All services operating

Select "Launch Instance" (midway down the page).



You are using the following Amazon EC2 resources in the US East (N. Virginia) region:

0 Running Instances	0 Elastic IPs
0 Volumes	0 Snapshots
0 Key Pairs	0 Load Balancers
0 Placement Groups	1 Security Group

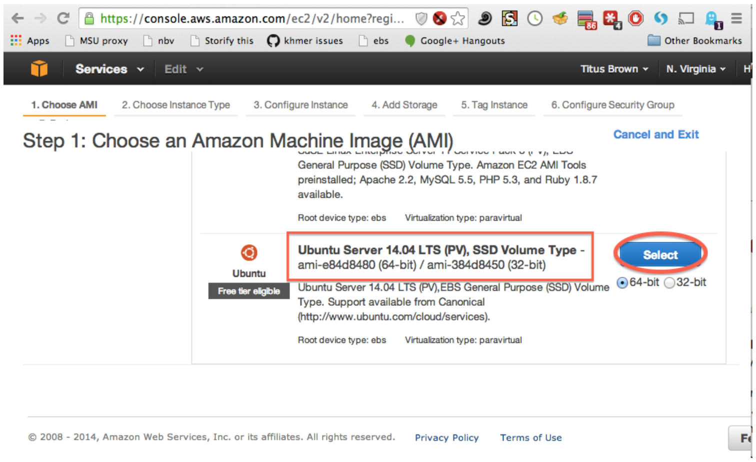
Easily deploy and operate applications - use Chef recipes, manage SSH users, and more. Try OpsWorks now.

To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.

Find **free software trial**

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Next, scroll down the list of operating system types until you find Ubuntu 14.04 LTS (PV) – it should be at the very bottom. Click 'select'. (See *Starting up a custom operating system* if you want to start up a custom operating system instead of Ubuntu 14.04.)



Step 1: Choose an Amazon Machine Image (AMI)

General Purpose (SSD) Volume Type. Amazon EC2 AMI Tools preinstalled; Apache 2.2, MySQL 5.5, PHP 5.3, and Ruby 1.8.7 available.

Root device type: ebs Virtualization type: paravirtual

Ubuntu
Free tier eligible

Ubuntu Server 14.04 LTS (PV), SSD Volume Type - ami-e84d8480 (64-bit) / ami-384d8450 (32-bit)

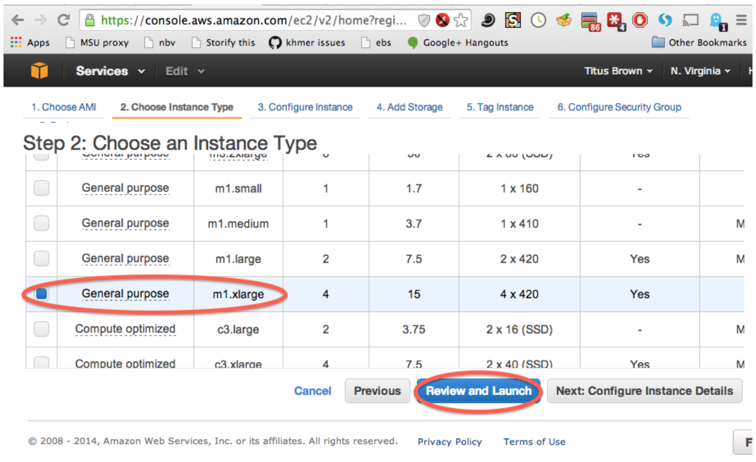
Ubuntu Server 14.04 LTS (PV),EBS General Purpose (SSD) Volume Type. Support available from Canonical (http://www.ubuntu.com/cloud/services).

Root device type: ebs Virtualization type: paravirtual

● 64-bit ○ 32-bit

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scroll down the list of instance types until you find “m1.xlarge”. Select the box to the left, and then click “Review and Launch.”

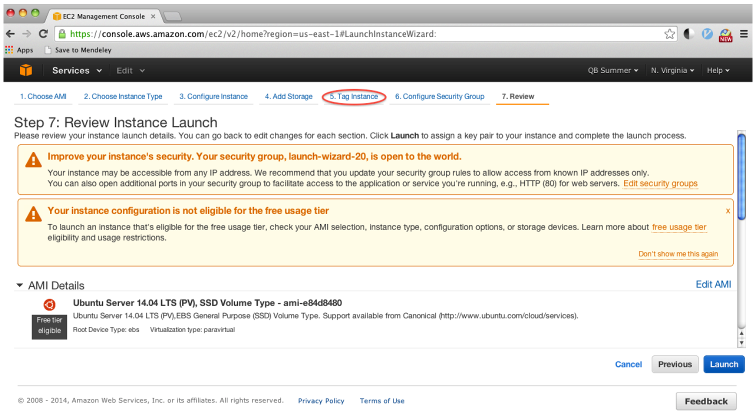


Step 2: Choose an Instance Type

□	General purpose	m1.small	1	1.7	1 x 160	-
□	General purpose	m1.medium	1	3.7	1 x 410	-
□	General purpose	m1.large	2	7.5	2 x 420	Yes
■	General purpose	m1.xlarge	4	15	4 x 420	Yes
□	Compute optimized	c3.large	2	3.75	2 x 16 (SSD)	-
□	Compute optimized	c3.xlarge	4	7.5	2 x 40 (SSD)	Yes

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Before the launch, we need to name our intances, so we can tell ours from others’. Click on “5.Tag Instance” at the top of that page.



Step 7: Review Instance Launch

Please review your instance launch details. You can go back to edit changes for each section. Click **Launch** to assign a key pair to your instance and complete the launch process.

Improve your instance's security. Your security group, launch-wizard-20, is open to the world.

Your instance may be accessible from any IP address. We recommend that you update your security group rules to allow access from known IP addresses only. You can also open additional ports in your security group to facilitate access to the application or service you're running, e.g., HTTP (80) for web servers.

Your instance configuration is not eligible for the free usage tier

To launch an instance that's eligible for the free usage tier, check your AMI selection, instance type, configuration options, or storage devices. Learn more about free usage tier eligibility and usage restrictions.

AMI Details

Free tier eligible

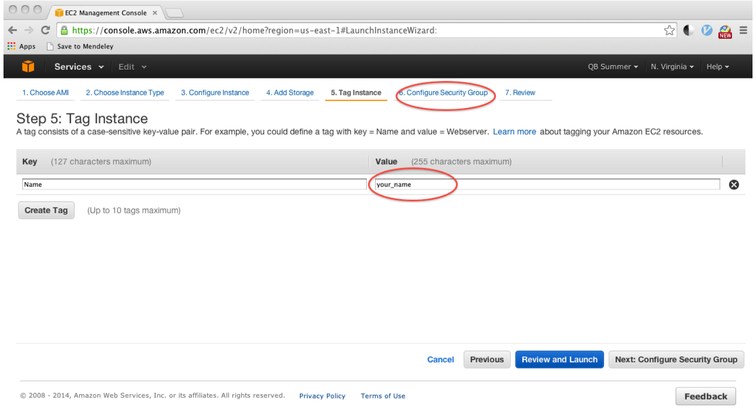
Ubuntu Server 14.04 LTS (PV), SSD Volume Type - ami-e84d8480

Ubuntu Server 14.04 LTS (PV),EBS General Purpose (SSD) Volume Type. Support available from Canonical (http://www.ubuntu.com/cloud/services).

Root Device Type: ebs Virtualization type: paravirtual

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Add an unique value at the “Value” column, e.g. your name. Then click on “6.Configure Security Group”.



Step 5: Tag Instance

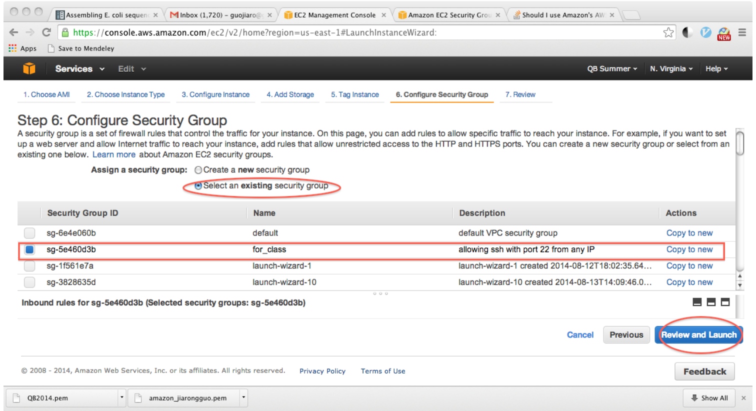
A tag consists of a case-sensitive key-value pair. For example, you could define a tag with key = Name and value = Webserver. Learn more about tagging your Amazon EC2 resources.

Key (127 characters maximum)	Value (255 characters maximum)
Name	your_name

(Up to 10 tags maximum)

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Choose “Select an existing security group”, select “for_class” under column “Name” and click “Review and Launch”:



Step 6: Configure Security Group

A security group is a set of firewall rules that control the traffic for your instance. On this page, you can add rules to allow specific traffic to reach your instance. For example, if you want to set up a web server and allow Internet traffic to reach your instance, add rules that allow unrestricted access to the HTTP and HTTPS ports. You can create a new security group or select from an existing one below. Learn more about Amazon EC2 security groups.

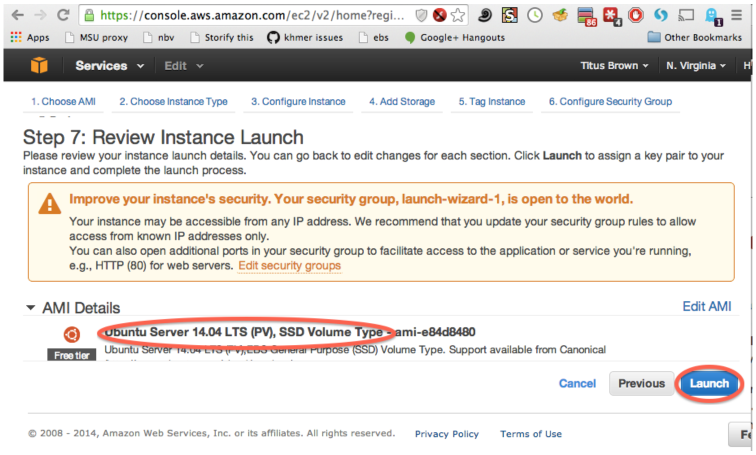
Assign a security group: ○ Create a **new** security group
● Select an **existing** security group

	Security Group ID	Name	Description	Actions
□	sg-6e4e060b	default	default VPC security group	Copy to new
■	sg-5e460d3b	for_class	allowing ssh with port 22 from any IP	Copy to new
□	sg-1f561e7a	launch-wizard-1	launch-wizard-1 created 2014-08-12T18:02:35.64...	Copy to new
□	sg-3828635d	launch-wizard-10	launch-wizard-10 created 2014-08-13T14:09:46.0...	Copy to new

Inbound rules for sg-5e460d3b (Selected security groups: sg-5e460d3b)

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Ignore the warning, double check that it says “Ubuntu 14.04 LTS (PV)” at AMI Details, and cick “Launch”.



Step 7: Review Instance Launch

Please review your instance launch details. You can go back to edit changes for each section. Click **Launch** to assign a key pair to your instance and complete the launch process.

Improve your instance's security. Your security group, launch-wizard-1, is open to the world.

Your instance may be accessible from any IP address. We recommend that you update your security group rules to allow access from known IP addresses only.
You can also open additional ports in your security group to facilitate access to the application or service you're running, e.g., HTTP (80) for web servers. Edit security groups

AMI Details

Ubuntu Server 14.04 LTS (PV), SSD Volume Type - ami-e84d8480

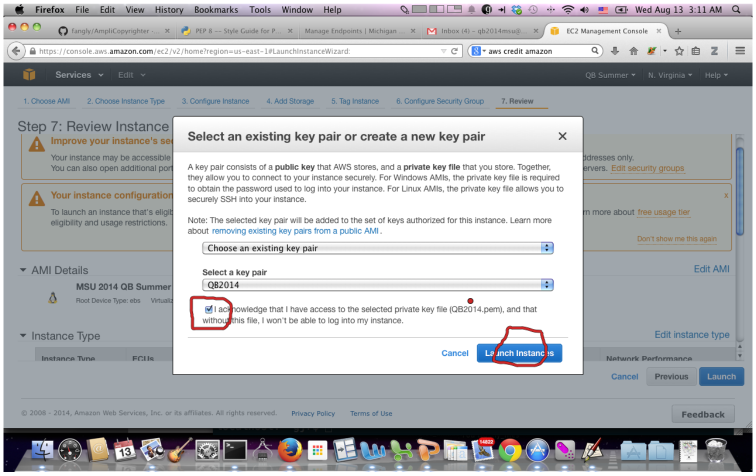
Ubuntu Server 14.04 LTS (PV),EBS General Purpose (SSD) Volume Type. Support available from Canonical

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Normally you will need to “create a new key pair” the first time through. However, for the purpose of easier management of our shared accout,

Select “Choose an existing key pair” and Select a key pair “QB2014”.

Select “Launch Instance.”



1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Tag Instance 6. Configure Security Group 7. Review

Step 7: Review Instance

Select an existing key pair or create a new key pair

A key pair consists of a **public key** that AWS stores, and a **private key file** that you store. Together, they allow you to connect to your instance securely. For Windows AMIs, the private key file is required to obtain the password used to log into your instance. For Linux AMIs, the private key file allows you to securely SSH into your instance.

Note: The selected key pair will be added to the set of keys authorized for this instance. Learn more about removing existing key pairs from a public AMI.

Choose an existing key pair

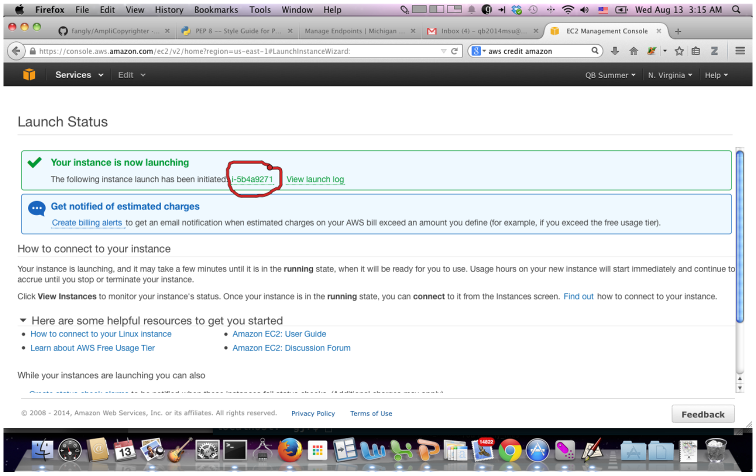
Select a key pair: QB2014

☑ I acknowledge that I have access to the selected private key file (QB2014.pem), and that without this file, I won't be able to log into my instance.

Cancel | Launch Instances

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

Click on the link of instance you just started.



Launch Status

Your instance is now launching

The following instance launch has been initiated: i-5b4a9271 View launch log

Get notified of estimated charges

Create billing alerts to get an email notification when estimated charges on your AWS bill exceed an amount you define (for example, if you exceed the free usage tier).

How to connect to your instance

Your instance is launching, and it may take a few minutes until it is in the **running** state, when it will be ready for you to use. Usage hours on your new instance will start immediately and continue to accrue until you stop or terminate your instance.

Click **View Instances** to monitor your instance's status. Once your instance is in the **running** state, you can **connect** to it from the Instances screen. Find out how to connect to your instance.

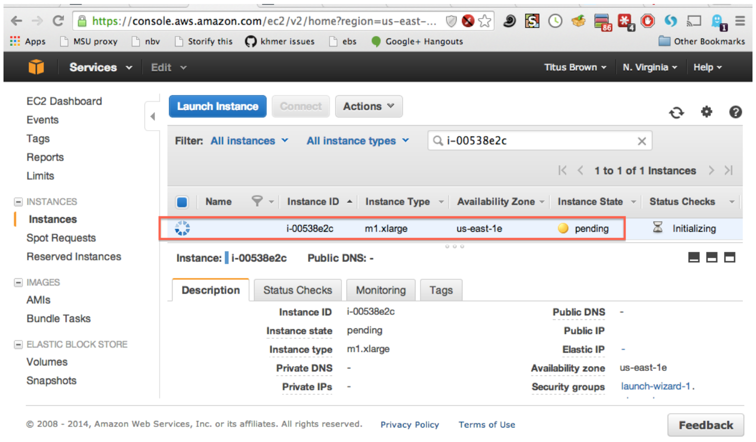
Here are some helpful resources to get you started

- How to connect to your Linux instance
- Learn about AWS Free Usage Tier
- Amazon EC2: User Guide
- Amazon EC2: Discussion Forum

While your instances are launching you can also

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

Then you should see a "pending" line in the menu.



Filter: All instances All instance types i-00538e2c

1 to 1 of 1 Instances

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks
	i-00538e2c	m1.xlarge	us-east-1e	pending	Initializing

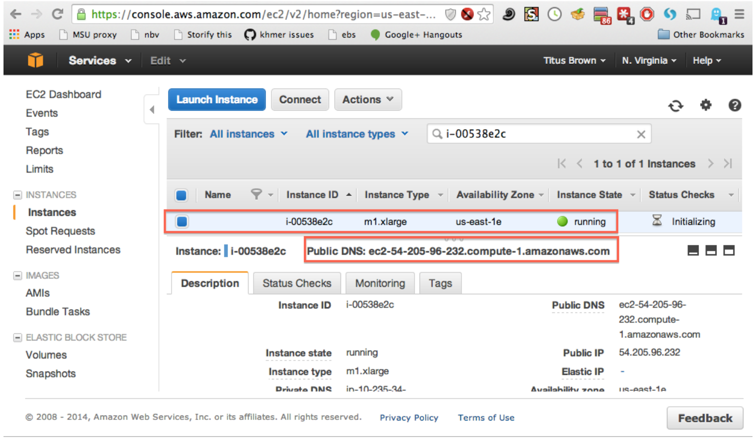
Instance: i-00538e2c Public DNS: -

Description | Status Checks | Monitoring | Tags

Instance ID	i-00538e2c	Public DNS	-
Instance state	pending	Public IP	
Instance type	m1.xlarge	Elastic IP	-
Private DNS	-	Availability zone	us-east-1e
Private IPs	-	Security groups	launch-wizard-1.

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

Wait until it turns green, then make a note of the "Public DNS" (we suggest copying and pasting it into a text notepad somewhere). This is your machine name, which you will need for logging in.



i-00538e2c

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks
	i-00538e2c	m1.xlarge	us-east-1e	● running	Initializing

Instance: i-00538e2c Public DNS: ec2-54-205-96-232.compute-1.amazonaws.com

Instance ID	i-00538e2c	Public DNS	ec2-54-205-96-232.compute-1.amazonaws.com
Instance state	running	Public IP	54.205.96.232
Instance type	m1.xlarge	Elastic IP	-
Private DNS	ip-10-235-34-	Availability zone	us-east-1e

Then, go to *Logging into your new instance "in the cloud" (Windows version)* or *Logging into your new instance "in the cloud" (Mac or Linux version)*

You might also want to read about *Terminating (shutting down) your EC2 instance*.

Logging into your new instance "in the cloud" (Mac or Linux version)

OK, so you've created a running computer. How do you get to it?

The main thing you'll need is the network name of your new computer. To retrieve this, go to the instance view and click on the instance, and find the "Public DNS". This is the public name of your computer on the Internet.

Copy this name, and connect to that computer with ssh under the username 'ubuntu', as follows:

First we need a private key file to log in; it is the .pem file we already created during the setup of amazon instance. For convience of managing the shared account, we want you all to use the same .pem file, which can be downloaded from:

http://lyorn.idyll.org/~gjr/public/QB2014/QB2014.pem

If your browser is firefox with default settings, it should be downloaded into your "Downloads" folder. Move it onto "Desktop".

If your browser is chrome with default settings, the above link will be shown as a page with text. right click and select "Save As". In the pop up window, change the directory to "Desktop" and the file name is still "QB2014.pem"

Other browsers should be similar to either of the above two case.

Next, start Terminal:

- in Applications... Utilities... for mac
- use shortcut "CTRL+ATL+T" for linux

and then type:

```
chmod og-rwx ~/Desktop/QB2014.pem
```

to set the permissions on the private key file to "closed to all evildoers".

Then type:

```
ssh -i ~/Desktop/QB2014.pem ubuntu@ec2-???-???-???-???.compute-1.amazonaws.com
```

(but you have to replace the stuff after the '@' sign with the name of the host).

Here, you're logging in as user 'ubuntu' to the machine 'ec2-174-129-122-189.compute-1.amazonaws.com' using the authentication key located in 'QB2014.pem' on your Desktop.

You should now see a text line that starts with something like `ubuntu@ip-10-235-34-223:~$`. You're in! Now type:

```
sudo bash
cd /root
```

to switch into superuser mode (see: http://xkcd.com/149/) and go to your home directory.

This is where the rest of the tutorials will start!

You might also want to read about *Terminating (shutting down) your EC2 instance*.

To log out, type:

```
exit
logout
```

or just close the window.

Logging into your new instance "in the cloud" (Windows version)

Download .pem file

First we need a private key file to log in; it is the .pem file we already created during the setup of amazon instance. For convience of managing the shared account, we want you all to use the same .pem file, which can be downloaded from:

http://lyorn.idyll.org/~gjr/public/QB2014/QB2014.pem

If your browser is firefox with default settings, it should be downloaded into your "Downloads" folder. Move it onto "Desktop".

If your browser is chrome with default settings, the above link will be shown as a page with text. right click and select "Save As". In the pop up window, change the directory to "Desktop" and the file name is still "QB2014.pem"

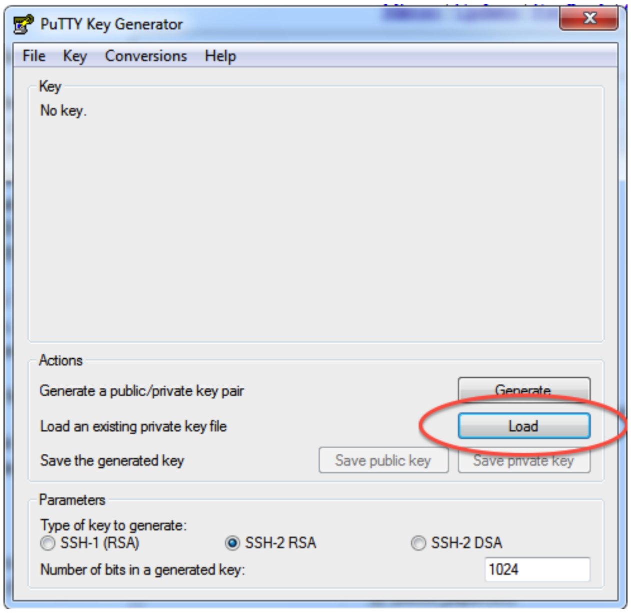
Other browsers should be similar to either of the above two case.

Download Putty and Puttygen from here: http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html

Generate a ppk file from your pem file

(You only need to do this once!)

Open puttygen; select "Load".



Key

No key.

Actions

Generate a public/private key pair

Load an existing private key file

Save the generated key

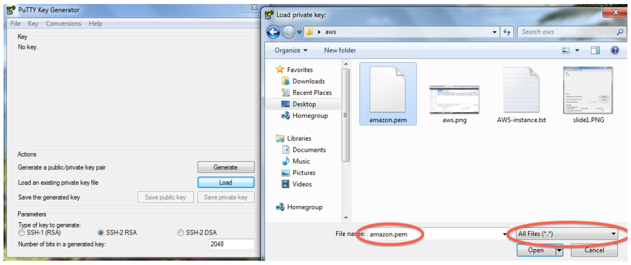
Parameters

Type of key to generate:

○ SSH-1 (RSA) ● SSH-2 RSA ○ SSH-2 DSA

Number of bits in a generated key: 1024

Find and load your '.pem' file; it's probably in your Downloads folder. Note, you have to select 'All files' on the bottom.



Key

No key.

Actions

Generate a public/private key pair

Load an existing private key file

Save the generated key

Parameters

Type of key to generate:

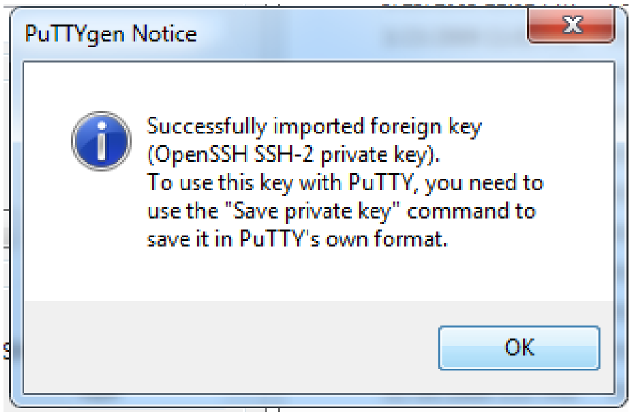
○ SSH-1 (RSA) ● SSH-2 RSA ○ SSH-2 DSA

Number of bits in a generated key: 2048

File name: amazon.pem

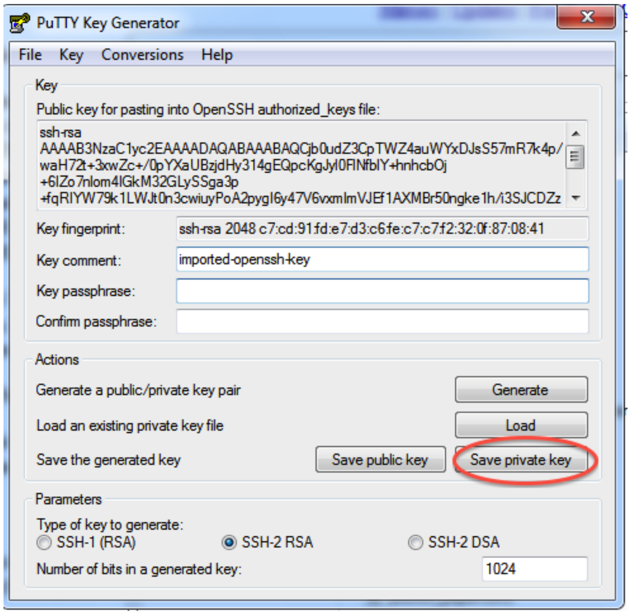
All Files (*.*)

Load it.



Successfully imported foreign key (OpenSSH SSH-2 private key).
To use this key with PuTTY, you need to use the "Save private key" command to save it in PuTTY's own format.

Now, "save private key". Put it somewhere easy to find.



Public key for pasting into OpenSSH authorized_keys file:

```
ssh-rsa
AAAAB3NzaC1yc2EAAAADAQABAAABAQCjb0udZ3CpTWZ4auWYxDJsS57mR7k4p/
waH72t+3xwZc+/0pYXaUBzjdHy314gEQpcKgJyl0FINfbIY+hnhcbOj
+6IZo7nlom4IGkM32GLySSga3p
+fqRIYW79k1LWJt0n3cwiuyPoA2pygl6y47V6vxmlmVJEf1AXMBr50ngke1h/i3SJCDZz
```

Key fingerprint: ssh-rsa 2048 c7:cd:91:fd:e7:d3:c6:fe:c7:c7:f2:32:0f:87:08:41

Key comment: imported-openssh-key

Key passphrase:

Confirm passphrase:

Generate a public/private key pair

Load an existing private key file

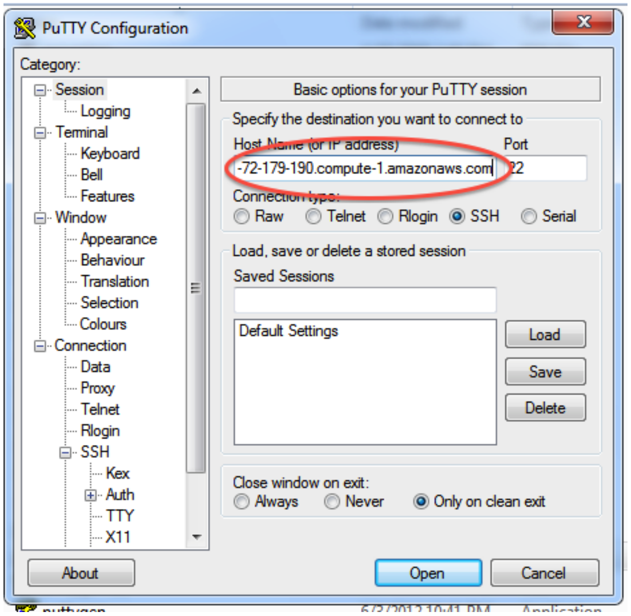
Save the generated key

Type of key to generate:
○ SSH-1 (RSA) ● SSH-2 RSA ○ SSH-2 DSA

Number of bits in a generated key: 1024

Logging into your EC2 instance with Putty

Open up putty, and enter your hostname into the Host Name box.



Specify the destination you want to connect to

Host Name (or IP address): -72-179-190.compute-1.amazonaws.com

Port: 22

Connection type:
○ Raw ○ Telnet ○ Rlogin ● SSH ○ Serial

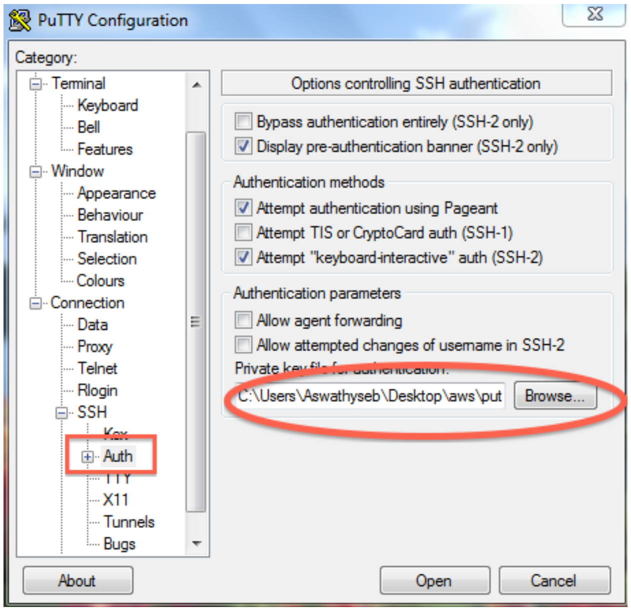
Load, save or delete a stored session

Saved Sessions

Default Settings

Close window on exit:
○ Always ○ Never ● Only on clean exit

Now, go find the 'SSH' section and enter your ppk file (generated above by puttygen). Then select 'Open'.



Options controlling SSH authentication

☐ Bypass authentication entirely (SSH-2 only)
☑ Display pre-authentication banner (SSH-2 only)

Authentication methods

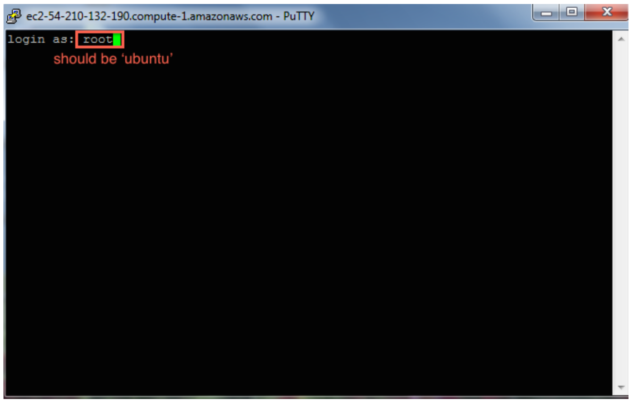
☑ Attempt authentication using Pageant
☐ Attempt TIS or CryptoCard auth (SSH-1)
☑ Attempt "keyboard-interactive" auth (SSH-2)

Authentication parameters

☐ Allow agent forwarding
☐ Allow attempted changes of username in SSH-2

Private key file for authentication:
C:\Users\Aswathyseb\Desktop\aws\put

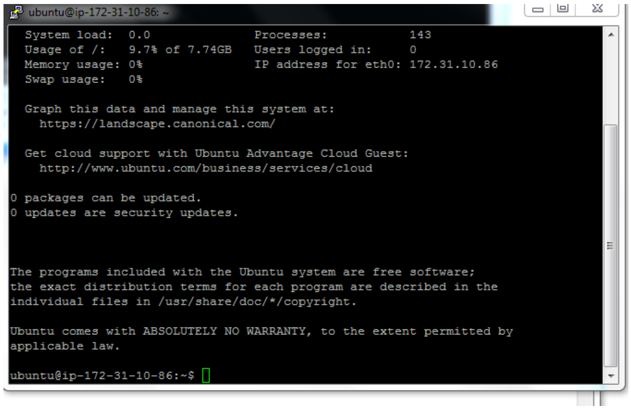
Log in as “ubuntu”.



```
login as: root
```

should be 'ubuntu'

Declare victory!



```
  System load:  0.0                Processes:           143
  Usage of /:   9.7% of 7.74GB   Users logged in:     0
  Memory usage: 0%                 IP address for eth0: 172.31.10.86
  Swap usage:   0%

  Graph this data and manage this system at:
    https://landscape.canonical.com/

  Get cloud support with Ubuntu Advantage Cloud Guest:
    http://www.ubuntu.com/business/services/cloud

0 packages can be updated.
0 updates are security updates.



The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

ubuntu@ip-172-31-10-86:~$
```

Here, you’re logging in as user ‘ubuntu’ to the machine ‘ec2-174-129-122-189.compute-1.amazonaws.com’ using the authentication key located in ‘amazon.pem’ on your Desktop.

You should now see a text line that starts with something like `ubuntu@ip-10-235-34-223:~$`. You’re in! Now type:

```
sudo bash
cd /root
```

to switch into superuser mode (see: http://xkcd.com/149/) and go to your home directory.

This is where the rest of the tutorials will start!

You might also want to read about *Terminating (shutting down) your EC2 instance*.

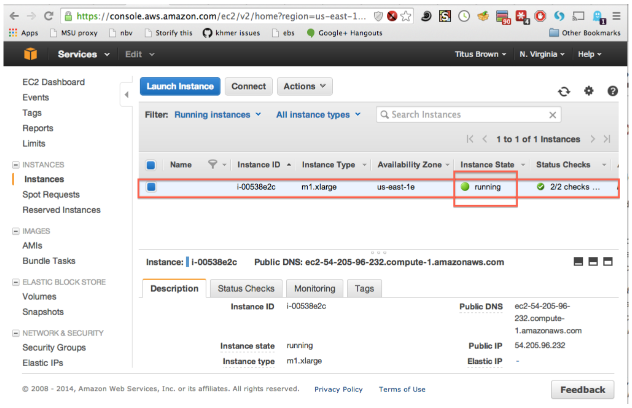
To log out, type:

```
exit
logout
```

or just close the window.

Terminating (shutting down) your EC2 instance

While your instance is running, Amazon will happily charge you on a per-hour basis – check out the pricing for more information. In general, you will want to shut down your instance when you're done with it; to do that, go to your EC2 console and find your running instances (in green).



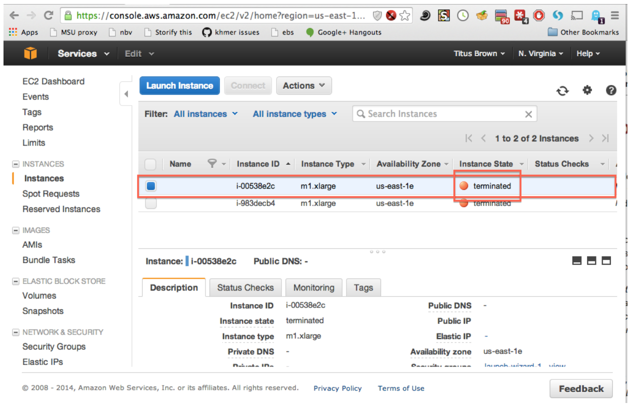
Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks
	i-00538e2c	m1.xlarge	us-east-1e	● running	✓ 2/2 checks ...

Instance: i-00538e2c Public DNS: ec2-54-205-96-232.compute-1.amazonaws.com

Instance ID	i-00538e2c	Public DNS	ec2-54-205-96-232.compute-1.amazonaws.com
Instance state	running	Public IP	54.205.96.232
Instance type	m1.xlarge	Elastic IP	-

Then, select one or all of them, and go to the 'Actions...' menu, and then select 'Terminate'. Agree.

After a minute or two, the console should show the instance as "terminated".



Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks
	i-00538e2c	m1.xlarge	us-east-1e	● terminated	
	i-983decb4	m1.xlarge	us-east-1e	● terminated	

Instance: i-00538e2c Public DNS: -

Instance ID	i-00538e2c	Public DNS	-
Instance state	terminated	Public IP	
Instance type	m1.xlarge	Elastic IP	-
Private DNS	-	Availability zone	us-east-1e

Starting up a custom operating system

The instructions in *Start up an EC2 instance* tell you how to start up a machine with Ubuntu Linux version 14.04 on it, but that machine comes with very little software installed. For anything where you are executing actual analyses, you're going to want to have a bunch of basic software installed.

Therefore, we make custom versions of Ubuntu available as well, that come with some software pre-installed.

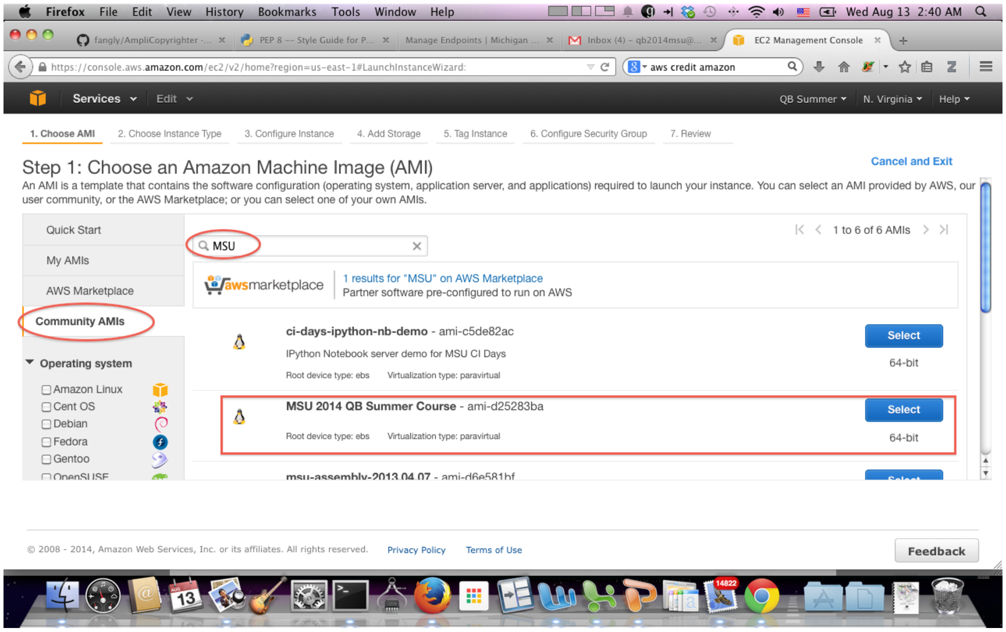
To boot these, go to EC2/Launch and select "Community AMIs" instead of the default "Quick Start";

Then type in the AMI number or name in the search box. We need to use:

ami-d25283ba

for this whole workshop. Or you can just type "MSU" in the search box and hit ENTER. **"MSU 2014 QB Summer Course"** is the one needed.

Then proceed with the rest of *Start up an EC2 instance*.



Step 1: Choose an Amazon Machine Image (AMI)

An AMI is a template that contains the software configuration (operating system, application server, and applications) required to launch your instance. You can select an AMI provided by AWS, our user community, or the AWS Marketplace; or you can select one of your own AMIs.

MSU

1 to 6 of 6 AMIs

1 results for "MSU" on AWS Marketplace
Partner software pre-configured to run on AWS

ci-days-ipython-nb-demo - ami-c5de82ac
IPython Notebook server demo for MSU CI Days
Root device type: ebs Virtualization type: paravirtual
64-bit

MSU 2014 QB Summer Course - ami-d25283ba
Root device type: ebs Virtualization type: paravirtual
64-bit

msu-assembly-2013.04.07 - ami-d6e581bf

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Variant calling

The goal of this tutorial is to show you the basics of variant calling using Samtools.

We'll be using data from one of Rich Lenski's LTEE papers, the one on the evolution of citrate consumption in LTEE.

Booting an Amazon AMI

Start up an Amazon computer (m1.large or m1.xlarge) using AMI ami-d25283ba (MSU 2014 QB Summer Course) (see *Start up an EC2 instance* and *Starting up a custom operating system*).

Log in the cloud machine from our local computer with Windows or from Mac OS X.

Logging in

Log in and type:

```
sudo bash
```

to change into superuser mode.

Software

Softwares are already installed in the ami loaded. Tools used in this tutorial are:

- bwa (mapping reads to reference genome)
- samtools (a fast and versatile tool for processing DNA alignment in SAM/BAM format.)

Download data

Data are already download in "/mnt" directory:

```
ls /mnt
```

to check the data files:

- assembly (directory of data for tomorrow's workshop)
- REL606.fa (reference genome)
- SRR098038.fastq.gz (reads from illumina sequencing)

Do the mapping

Now let's map all of the reads to the reference. Start by indexing the reference genome:

```
cd /mnt

bwa index REL606.fa
```

Now, do the mapping of the raw reads to the reference genome:

```
bwa aln REL606.fa SRR098038.fastq.gz  > SRR098038.sai
```

Make a SAM file (this would be done with 'sampe' if these were paired-end reads):

```
bwa samse REL606.fa SRR098038.sai SRR098038.fastq.gz > SRR098038.sam
```

This file contains all of the information about where each read hits on the reference.

Next, index the reference genome with samtools:

```
samtools faidx REL606.fa
```

Convert the SAM into a BAM file:

```
samtools import REL606.fa.fai SRR098038.sam SRR098038.bam
```

Sort the BAM file:

```
samtools sort SRR098038.bam SRR098038.sorted
```

And index the sorted BAM file:

```
samtools index SRR098038.sorted.bam
```

Visualizing alignments

‘samtools tview’ is a text interface that you use from the command line; run it like so:

```
samtools tview SRR098038.sorted.bam REL606.fa
```

The ‘.’s are places where the reads align perfectly in the forward direction, and the ‘,’s are places where the reads align perfectly in the reverse direction. Mismatches are indicated as A, T, C, G, etc.

You can scroll around using left and right arrows; to go to a specific coordinate, use ‘g’ and then type in the contig name and the position. For example, type ‘g’ and then ‘rel606:553093<ENTER>’ to go to position 553093 in the BAM file.

Use ‘q’ to quit.

Counting alignments

This command:

```
samtools view -c -f 4 SRR098038.bam
```

will count how many reads DID NOT align to the reference (214518).

This command:

```
samtools view -c -F 4 SRR098038.bam
```

will count how many reads DID align to the reference (6832113).

And this command:

```
gunzip -c SRR098038.fastq.gz | wc
```

will tell you how many lines there are in the FASTQ file (28186524). Reminder: there are four lines for each sequence.

Calling SNPs

You can use samtools to call SNPs like so:

```
samtools mpileup -uD -f REL606.fa SRR098038.sorted.bam | bcftools view -bvcg - > SRR098038.raw.bcf
```

(See the ‘mpileup’ docs here.)

Now convert the BCF into VCF:

```
bcftools view SRR098038.raw.bcf > SRR098038.vcf
```

You can check out the VCF file by using ‘tail’ to look at the bottom:

```
tail *.vcf
```

Each variant call line consists of the chromosome name (for E. coli REL606, there’s only one chromosome - rel606); the position within the reference; an ID (here always ‘.’); the reference call; the variant call; and a bunch of additional information.

Again, you can use ‘samtools tview’ and then type (for example) ‘g’ ‘rel606:4616538’ to go visit one of the positions. The format for the address to go to with ‘g’ is ‘chr:position’.

You can read more about the VCF file format here.

Questions/discussion items

Why so many steps?

Assembling E. coli sequences with Velvet

The goal of this tutorial is to show you the basics of assembly using the Velvet assembler.

We’ll be using data from Efficient de novo assembly of single-cell bacterial genomes from short-read data sets, Chitsaz et al., 2011.

Booting an Amazon AMI

Start up an Amazon computer (m1.large or m1.xlarge) using AMI ami-d25283ba (see *Start up an EC2 instance* and *Starting up a custom operating system*).

Log in with Windows or from Mac OS X.

Logging in

Log in and type:

```
sudo bash
```

to change into superuser mode.

Packages to install

Softwares and packages are already installed:

- velvet (An assmbler)
- khmer (A package for preprocess and assemble large data)
- quast (A tools for evaluating assembly with known genomes)

Data

Data is already downloaded in /mnt/assembly:

```
ls /mnt/assembly
```

to see data in "/mnt/assembly":

- ecoli_ref-5m-trim.fastq.gz: quality trimmed pair end data sets from E. coli genome sequence data.
- ecoli-reads-5m-dn-paired.fa.gz: data set is a specially processed data set using digital normalization that will assemble quickly.

Running an assembly

Go inside the data directory:

```
cd /mnt/assembly
```

Now... assemble the small, fast data sets, using the Velvet assembler. Here we will set the required parameter k=21:

```
velveth ecoli.21 21 -shortPaired -fasta.gz ecoli-reads-5m-dn-paired.fa.gz
velvetg ecoli.21 -exp_cov auto
```

Check out the stats for the assembled contigs for a cutoff of 1000:

```
python /usr/local/share/khmer/sandbox/assemstats3.py 1000 ecoli.*/contigs.fa
```

Also try assembling with k=23 and k=25:

```
velveth ecoli.23 23 -shortPaired -fasta.gz ecoli-reads-5m-dn-paired.fa.gz
velvetg ecoli.23 -exp_cov auto

velveth ecoli.25 25 -shortPaired -fasta.gz ecoli-reads-5m-dn-paired.fa.gz
velvetg ecoli.25 -exp_cov auto
```

Now check out the stats for the assembled contigs for a cutoff of 1000:

```
python /usr/local/share/khmer/sandbox/assemstats3.py 1000 ecoli.*/contigs.fa
```

(Also read: What does k control in de Bruijn graph assemblers?.)

Comparing and evaluating assemblies - QUAST

Run QUAST to mapping contigs (ecoli.23/contigs.fa, and ecoli.25/contigs.fa) reference genome (ecoliMG1655.fa):

```
gunzip ecoliMG1655.fa.gz
/root/quast-2.3/quast.py -R ecoliMG1655.fa ecoli.*/contigs.fa
```

Note that here we're looking at *all* the assemblies we've generated.

Now look at the results:

```
more quast_results/latest/report.txt
```

The first bits to look at are Genome fraction (%) and # misassembled contigs, I think.

Searching assemblies – BLAST

Build BLAST databases for the assemblies you've done:

```
cd /mnt/assembly

for i in 21 23 25
do
   extract-long-sequences.py -o ecoli-$i.fa -l 500 ecoli.$i/contigs.fa
   formatdb -i ecoli-$i.fa -o T -p F
done
```

BLAST is already installed. Let's search for a specific gene (CRP, a transcription regulator).

The commandline for BLAST is:

```
blastall -i crp.fa -d ecoli-21.fa -p tblastn -b 1 -v 1
```

Questions and Discussion Points

Why do we use a lower cutoff of 1kb for the assemstats3 links, above? Why not 0?

CHAPTER 2

Indices and tables

- genindex
- modindex
- search