

NGS Data Analysis and Galaxy

University of Cape Town
Cape Town, South Africa
21-25 October 2013

Dave Clements, Emory University
<http://galaxyproject.org/>

Gerrit Botha
Computational Biology
University of Cape Town
<http://cbio.uct.ac.za/>



UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

South Africa
Galaxy
Workshop Tour

This Week

Monday	Welcome, Project Intro, Basic Galaxy Usage NGS QualityControl
Tuesday	RNA-Seq - Mapping and Transcript Prediction RNA-Seq: Differential expression and Alternative Pipelines; SNP & Variant Analysis
Wednesday	SNP & Variant Analysis Chip-Seq Analysis
Thursday	Genome Assembly Install your own Galaxy on Amazon Cloud
Friday	Customizing Galaxy, Galaxy Tool Shed, and Wrapping Tools for Galaxy

Friday Agenda

- 9:00 **Welcome and Questions**
- 9:15 **Intro to the Linux command line**
- 10:15 **Galaxy Revealed! Part I**
- 11:00 Break
- 11:30 Galaxy Revealed! Part II
- 13:00 Lunch
- 14:00 Galaxy Tool Wrapping Part I
- 15:30 Break
- 16:00 Galaxy Tool Wrapping Part II
- 17:00 Done

Create a new Cloud Instance

<http://bit.ly/??????>

Save your PEM file!

Create an account on the server
and make it an admin through the
CloudMan interface.

Using the shell from Windows: PuTTY

Windows Desktop → Healthlab →
metagenomics_course_nov12

puttygen:

Convert pem to a format PuTTY can use

putty:

Connect to the command line

Using PuTTY with AWS

<http://bit.ly/AWSputty>

SSH Magic (not using PuTTY)

```
$ ssh -i whatever.pem \  
    ubuntu@ec2-xxx-xxx-xxx-xxx.region.compute.amazonaws.com  
The authenticity of host 'ec2-xxx-xxx-xxx-  
xxx.region.compute.amazonaws.com (xxx-xxx-xxx-xxx)' can't be  
established.  
RSA key fingerprint is  
xx:xx:xx:xx:xx:xx:xx:xx:xx:xx:xx:xx:xx:xx:xx:xx.  
Are you sure you want to continue connecting (yes/no)? yes  
Warning: Permanently added 'ec2-xxx-xxx-xxx-  
xxx.region.compute.amazonaws.com' (RSA) to the list of known  
hosts.  
Welcome to Ubuntu 12.04.3 LTS (GNU/Linux 3.2.0-53-virtual  
x86_64)  
...  
ubuntu@ip-xxx-xxx-xxx-xxx:~$
```

And now you are in the command line shell of your
Galaxy instance

A word on users

```
ubuntu@ip-xxx-xxx-xxx-xxx:~$
```

```
ubuntu$
```

```
$
```

ubuntu at front of the **prompt** means you are now logged in as the ubuntu user. We'll spend all of our time as one of these:

ubuntu:

Has the ability to *invoke superpowers and to become other users*

galaxy:

Owens the Galaxy installation and files. Has no superpowers.

Now what?

```
ubuntu@ip-xxx-xxx-xxx-xxx:~$
```

<http://overapi.com/linux/>

Now what?

<http://overapi.com/linux/>

Where am I? (In my *home* directory)

Create a new directory named Messy

Make Messy my *current* directory

Create a new directory under Messy

Create a 2nd new directory under Messy

What's in Messy so far?

What's in Messy, exactly?

Make workspace my current directory

Where am I? (In ~/Messy/workspace)

What's here? (nothing)

What's in my *parent* directory, Messy?

Change to workspace's parent directory

Where am I? (In ~/Messy)

```
$ pwd
$ mkdir Messy
$ cd Messy
$ mkdir MessySubDir
$ mkdir workspace
$ ls
$ ls -l
$ cd workspace
$ pwd
$ ls
$ ls ..
$ cd ..
$ pwd
```

A word on *sudo*

```
ubuntu$ a command ubuntu does not have permission for
Permission denied.
ubuntu$ sudo same command
ubuntu$

ubuntu$ sudo su -l galaxy
galaxy$ sudo anything
Permission denied.
galaxy$ exit
ubuntu$
```

The `sudo` command allows certain users to do things normal users can't do, including become other users.

ubuntu can run `sudo`.

galaxy cannot run `sudo`.

Where things are

```
$ cd /mnt
$ ls
cm galaxy galaxyData galaxyIndices lost+found transient_nfs
$
```

cm

CloudMan details

galaxy

The running Galaxy instance. We'll spend most of our time here

galaxyData

Empty until something happens

galaxyIndices

All the reference data that comes packaged with Galaxy on the cloud

The Universe!

```
$ sudo su -l galaxy
$ cd /mnt/galaxy/galaxy-app
$ less universe_wsgi.ini
$ pico universe_wsgi.ini
$
```

universe_wsgi.ini

Galaxy instance configuration details

Add a brand to your server.

In the [app:main] section, add

```
brand = My Cool Brand
```

Restart Galaxy through CloudMan admin page and then refresh your main Galaxy page.

Friday Agenda

- 8:30 **Welcome and Questions**
- 8:45 **Intro to the Linux command line**
- 9:45 **Galaxy Revealed! Part I**
- 10:30 **Break**
- 11:00 **Galaxy Revealed! Part II**
- 12:30 **Lunch**
- 13:30 **Galaxy Tool Wrapping Part I**
- 15:00 **Break**
- 15:30 **Galaxy Tool Wrapping Part II**
- 16:30 **Done**

Friday Agenda

- 8:30 **Welcome and Questions**
- 8:45 **Intro to the Linux command line**
- 9:45 **Galaxy Revealed! Part I**
- 10:30 **Break**
- 11:00 **Galaxy Revealed! Part II**
- 12:30 **Lunch**
- 13:30 **Galaxy Tool Wrapping Part I**
- 15:00 **Break**
- 15:30 **Galaxy Tool Wrapping Part II**
- 16:30 **Done**

Add a reference genome: 1st step

Define the DB KEY:

```
$ cd /mnt/galaxy/galaxy-app/tool-data/shared/ucsc/  
$ pico builds.txt  
$ pico universe_wsgi.ini  
$
```

Add

```
milliV      Milli Vanilli (milliV)
```

as last line. That is a **TAB, not spaces**, between columns 1 and 2.

Save the file.

Add a reference genome: Data Managers

<http://bit.ly/gxyDataManagers>

<http://bit.ly/gccDataManagers>

Cut and paste and follow directions on HowTo page.

Use defaults

```
galaxy$ cd /mnt/galaxy/galaxy-app/tools
$ mkdir data_manager
$ cd data_manager
$ pico fetch_genome_all_fasta.xml
$ pico data_manager_fetch_genome_all_fasta.py
$ cd ../..
$ pico data_manager_conf.xml
$ pico universe_wsgi.ini
```

Restart Galaxy

Add a reference genome: Data Managers

On your **Cloud[123].galaxyproject.org**

Shared Data → Published Histories → Milli v 101

Click on [m.vannielli.sequence.fasta](#)

Hover over the diskette icon.

Copy the URL

On your **own cloud instance**

Analyze Data → Get Data

→ Upload File

Add a reference genome: Data Managers

On your cloud instance, using the Galaxy GUI

Admin → Data → Manage Local Data

Add a reference genome: Data Managers

On your cloud instance, using the Galaxy GUI

Admin → Data → Manage Local Data

Where is it? Let's look at the log.

```
galaxy$ cd /mnt/galaxy/galaxy-app/  
$ less main.log  
  
galaxy.tools.data_manager.manager ERROR 2013-10-18  
05:53:55,807 There was an error parsing your Data Manager  
config file "./data_manager_conf.xml": mismatched tag: line  
19, column 6  
  
$ nl data_manager_conf.xml  
$ nano data_manager_conf.xml
```

Restart Galaxy

Add a reference genome: Data Managers

On your cloud instance, using the Galaxy GUI

Admin → Data → Manage Local Data

→ Access Data Managers

→ Reference Genomes

Source → History or URL

Add a reference genome: Data Managers

Cool!

but...

NGS: Mapping → Bowtie

Ain't no Milli V.

Add a reference genome: Data Managers

Bowtie has its own indexes for genomes.
Lots of tools require preprocessing steps.

Need to tell Bowtie about Milli V.

Admin → Search and Browse Tool Sheds
Select Galaxy Test Tool Shed

Search for bowtie

Select data_manager_bowtie_index_builder

Click Install to Galaxy

Add a reference genome: Data Managers

Repository installation is not possible due to an invalid Galaxy URL: None. You may need to enable cookies in your browser.

Well, how did we do it back in 2012?

```
galaxy$ cd /mnt/galaxy/galaxy-app/tool-data/milliV/seq
$ ls
$ ../../../../../../tools/bowtie/default/bin/bowtie-build -f milliV1.fa milliV1
$ mkdir ../bowtie
$ mv milliV1.*ebwt ../bowtie/
$ pico /mnt/galaxyIndices/galaxy/tool-data/bowtie_indices.loc

milliV1      milliV1      milliV1: Milli Vanilli      /mnt/galaxy/galaxy-app/tool-
data/milliV/bowtie/milliV1
```

Make sure those are TABs, not spaces.

Add a Tool:

Admin → Search and Browse Tool Sheds

Select Galaxy Main Tool Shed

Search for Venn

Select venn_list

Click Install to Galaxy

Define your own tools: R

[http://wiki.galaxyproject.org/News/
RGalaxyWrapRFunctionsAsTools](http://wiki.galaxyproject.org/News/RGalaxyWrapRFunctionsAsTools)

[http://bioconductor.org/packages/release/bioc/html/
RGalaxy.html](http://bioconductor.org/packages/release/bioc/html/RGalaxy.html)

[http://toolshed.g2.bx.psu.edu/view/nikhil-joshi/
deseq_and_sam2counts](http://toolshed.g2.bx.psu.edu/view/nikhil-joshi/deseq_and_sam2counts)

Create our own tool

[http://wiki.galaxyproject.org/Events/GCC2013/
TrainingDay](http://wiki.galaxyproject.org/Events/GCC2013/TrainingDay)

Thanks



Dave Clements

**Galaxy Project
Emory University**

clements@galaxyproject.org