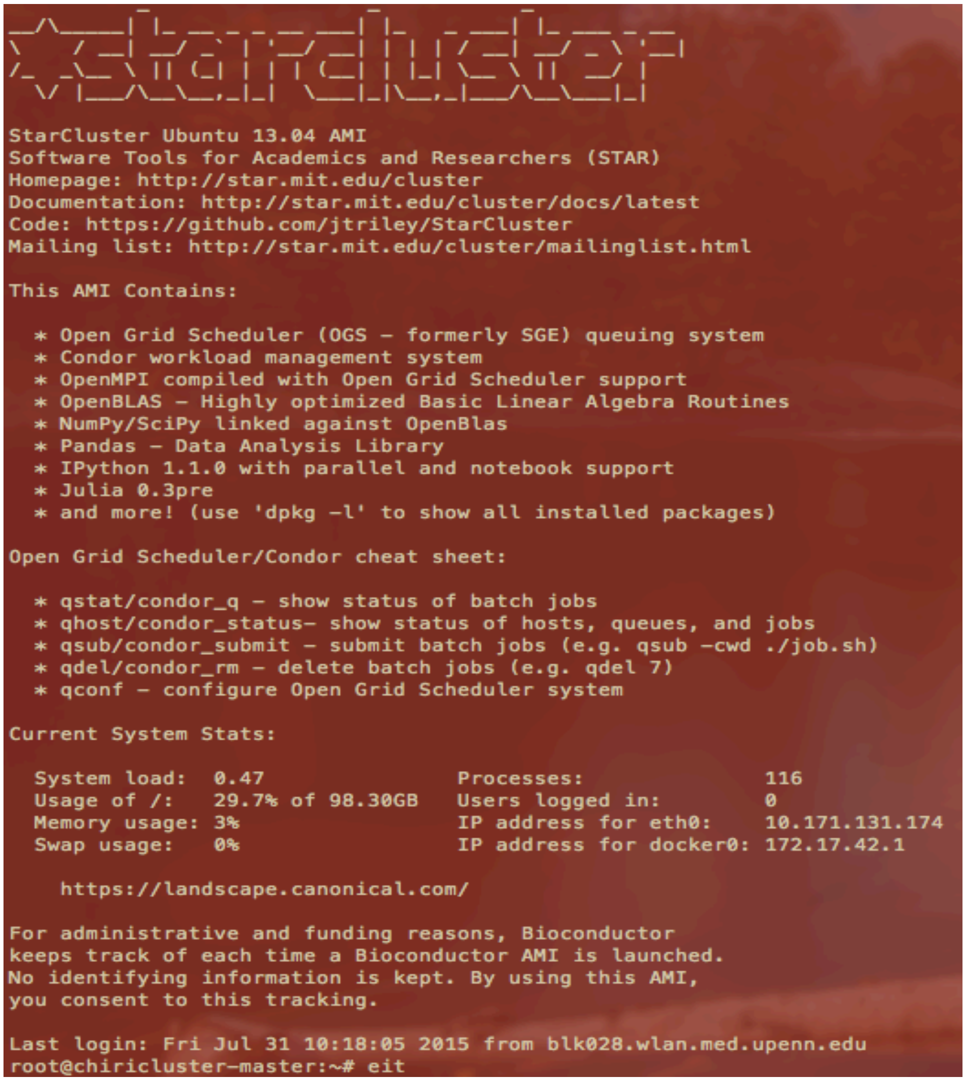
Using Amazon EC2 Cluster

Starcluster Method

1. To begin using the Amazon Cluster, it is first necessary to log on to the master node of the cluster via the command Terminal
   - First, move the keypair assigned to cluster into a folder on a desktop
     - For ease sake, create an empty folder on the desktop and name it "key"
   - Open the command terminal, set the directory to the folder of the key | type:
     - cd desktop
     - cd key
   - Now modify the permission of the key | type:
     - chmod 400 amazonpair.pem
   - To log into the cluster | type:
     - ssh –i amazonpair.pem ubuntu@public DNS name
       - Example of login to cluster with the current DNS name
         - ssh –i amazonpair.pem ubuntu@ec2-54-147-214-86.compute-1.amazonaws.com



```
                 _                 _           _
__/\_____| |_ __ _ _ __ ___| |_   _ ___| |_ ___ _ __
\    / __| __/ _` | '__/ __| | | | / __| __/ _ \ '__|
/_  _\__ \ || (_| | | | (__| | |_| \__ \ ||  __/ |
  \/ |___/\__\__,_|_|  \___|_|\__,_|___/\__\___|_|

StarCluster Ubuntu 13.04 AMI
Software Tools for Academics and Researchers (STAR)
Homepage: http://star.mit.edu/cluster
Documentation: http://star.mit.edu/cluster/docs/latest
Code: https://github.com/jtriley/StarCluster
Mailing list: http://star.mit.edu/cluster/mailinglist.html

This AMI Contains:

  * Open Grid Scheduler (OGS - formerly SGE) queuing system
  * Condor workload management system
  * OpenMPI compiled with Open Grid Scheduler support
  * OpenBLAS - Highly optimized Basic Linear Algebra Routines
  * NumPy/SciPy linked against OpenBlas
  * Pandas - Data Analysis Library
  * IPython 1.1.0 with parallel and notebook support
  * Julia 0.3pre
  * and more! (use 'dpkg -l' to show all installed packages)

Open Grid Scheduler/Condor cheat sheet:

  * qstat/condor_q - show status of batch jobs
  * qhost/condor_status- show status of hosts, queues, and jobs
  * qsub/condor_submit - submit batch jobs (e.g. qsub -cwd ./job.sh)
  * qdel/condor_rm - delete batch jobs (e.g. qdel 7)
  * qconf - configure Open Grid Scheduler system

Current System Stats:

  System load:  0.47               Processes:              116
  Usage of /:   29.7% of 98.30GB   Users logged in:        0
  Memory usage: 3%                 IP address for eth0:    10.171.131.174
  Swap usage:   0%                 IP address for docker0: 172.17.42.1

    https://landscape.canonical.com/

For administrative and funding reasons, Bioconductor
keeps track of each time a Bioconductor AMI is launched.
No identifying information is kept. By using this AMI,
you consent to this tracking.

Last login: Fri Jul 31 10:18:05 2015 from blk028.wlan.med.upenn.edu
root@chiricluster-master:~# eit
```

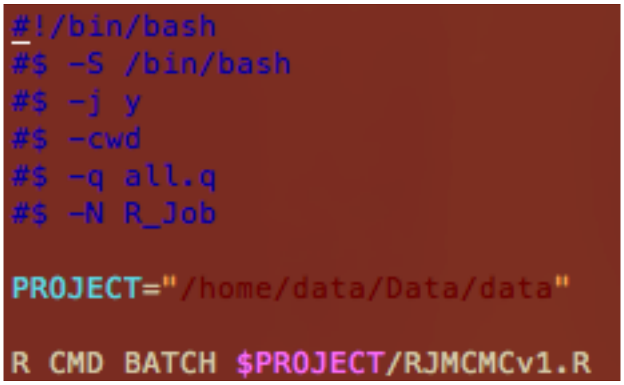
   - You should now be logged into the master node of the cluster. If it takes a long amount of time to do so and a timeout message regarding ssh port 22 is received, then ensure that the internet connection being used is functional

Using Amazon EC2 Cluster

Starcluster Method

2. Once you are inside of the cluster environment, you can now move data to and fro the cluster and submit R jobs
    - It is important to note that the cluster uses an Elastic Block Storage (EBS) system to save and share files among the machines
        - This ensures that when a cluster or a node is terminated that the information utilized and outputted is still available
        - Its similar to an external hard drive
        - If you are in the /home/ubuntu directory, then you are within the local machine and all information placed here is temporary and will be deleted upon termination of the cluster
        - If you are in the outer /home/data directory, then you are within the created EBS storage system and all saved storage is permanent
    - Thus to ensure that you are in the directory of the EBS system move into the outer "home" directory | type:
        - cd /home/data
    - Now can make directories within this environment and it would be permanent
    - To list the directories in this directory | type:
        - ls
    - To make a directory | type:
        - mkdir [name of folder]
        - cd [name of folder] ##to move into that folder
    - To move data into the cluster, open another tap in the terminal (found at the top under "shell")
        - Move to directory with keypair
            - ➢ cd desktop
            - ➢ cd key
        - Moving a single file to the cluster | type:
            - ➢ scp –i amazonpair.pem /path/to/file/in/directory/file.R ubuntu@publicDNSname:/path/desired/in/the/cluster
                - Example: scp –i amazonpair.pem /Users/patrickemedom/Desktop/Levy_lab/Jewell/SensAnalysis/RJMCMCv1.R ubuntu@ec2-54-205-37-80.compute-1.amazonaws.com:/home/data/Jewell
        - Moving an entire directory into the cluster (only difference is adding the recursive command, -r) | type:
            - ➢ scp –r –i amazonpair.pem /path/to/directory/ ubuntu@publicDNSname:/path/desired/in/the/cluster
    - To move data from the cluster to local computer
        - Open new tap in terminal and move to location of keypair.pem
            - ➢ cd desktop
            - ➢ cd key
        - Moving a single file to computer | type:
            - ➢ scp –i amazonpair.pem ubuntu@ec2-54-147-214-86.compute-1.amazonaws.com:/path/in/cluster/to/file/Results1.csv /path/to/desired/location/on/computer/
            - ➢ example: scp –r –i amazonpair.pem scp -i amazonpair.pem ubuntu@ec2-54-147-214-86.compute-1.amazonaws.com:/home/data/Data/data/Results1.csv /Users/patrickemedom/Desktop/Levy_Lab/Jewell
        - Moving an entire directory to local computer | type:

- scp –r –i amazonpair.pem ubuntu@ec2-54-147-214-86.compute-1.amazonaws.com:/path/in/cluster/to/directory/ /path/to/desired/location/on/computer/

3. Submitting R jobs to the Amazon Cluster
    - Luckily the cluster is equipped with a Sun Grid Engine queueing system, which makes submitting jobs to the nodes of the cluster fairly simple
        - In order to submit R jobs two things are needed, the desired R script and the wrapper.sh script
        - First move the R script and the wrapper script in the same directory in the cluster
            - The wrapper script can be found in the directory /home/data
            - To copy the wrapper to desired location, type:
                - cp wrapper.sh /path/to/desired/location
        - Edit the wrapper script to read R script
            - Use the Vi script editor
                - vi wrapper.sh
            - Tips to using Vi script editor – while in Vi type
                - i #insert text before cursor, until <Esc> hit
                - u #undo whatever you just did
                - x #delete single character under cursor
                - :x #quit vi, saving the latest edit under the original file name
                - :q! #quit vi even though latest changes have not been saved for this vi call
            - Wrapper.sh
                - 

```
#!/bin/bash
#$ -S /bin/bash
#$ -j y
#$ -cwd
#$ -q all.q
#$ -N R_Job

PROJECT="/home/data/Data/data"

R CMD BATCH $PROJECT/RJMCMCv1.R
```

                - Under project name the path to the directory of the R script
                - After $PROJECT/(insert name of R script)
                - Type: :x (to save and exit vi)
    - Now you can submit the R job to the cluster | type:
        - qsub wrapper.sh
    - To view the status of the R Job | type:
        - qstat
            - Active jobs will be present here, while jobs that have either been completed or have failed to be submitted will not.
        - qhost
            - To view the cpu load of each node, which gives a general idea of which nodes are running the jobs
    - Tips
        - It is important to lot that the cluster is set so that only one job can run on one node

- If a job fails to be submitted ensure that the path in the wrapper script correct and the spelling of the R script is correct
- You will still need to install packages to the nodes
    - I would advise making a R script containing all of the packages required, then edit the wrapper to read the R script and place jobs ("qsub wrapper.sh") until all the nodes are occupied

4. Exiting the cluster | type:
    - exit