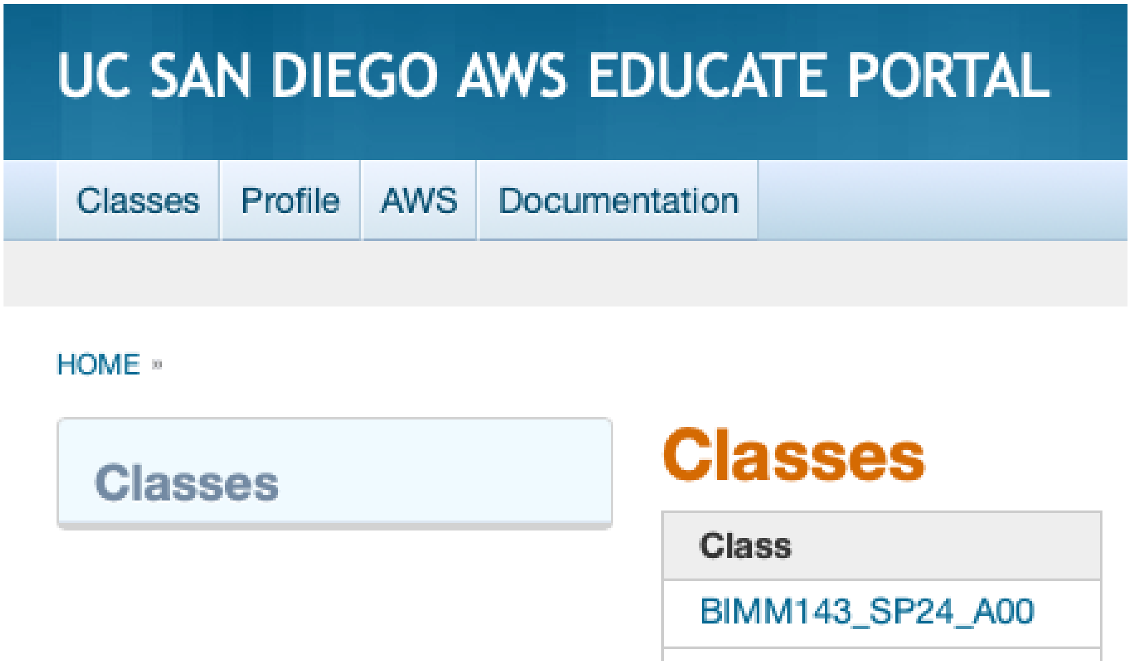
17_Lab17_Analyzing_sequencing_data_in_the_cloud_[Extra_credit]

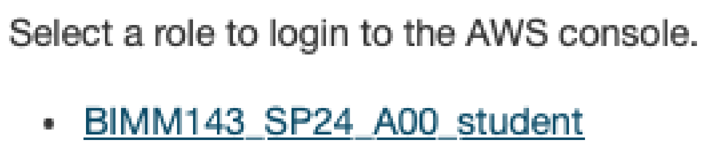
June 5, 2024

1 Login

https://awsed.ucsd.edu/

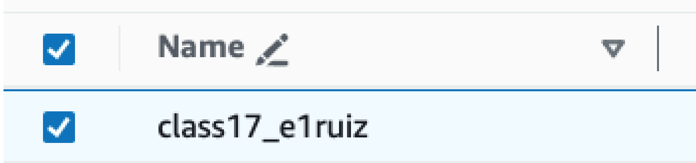


Class
BIMM143_SP24_A00



Select a role to login to the AWS console.

- BIMM143_SP24_A00_student



☑	Name
☑	class17_e1ruiz

2 Terminal Commands to setup UNIX

```
(base) er@er-MacBook-Pro BIMM143 % cd Class\ 17
(base) er@er-MacBook-Pro Class 17 % ls
bimm143_e1ruiz.pem
(base) edwinruiz@Edwins-MacBook-Pro Class 17 % chmod 400 "bimm143_e1ruiz.pem"
(base) edwinruiz@Edwins-MacBook-Pro Class 17 % ssh -i "bimm143_e1ruiz.pem" ubuntu@ec2-34-219-1
```

```
The authenticity of host 'ec2-34-219-168-96.us-west-2.compute.amazonaws.com (34.219.168.96)' ca
ED25519 key fingerprint is SHA256:Ptz4XmSuaNkt65ielCTcPlH6CyDQAYPT/t+65mRFUTc.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-34-219-168-96.us-west-2.compute.amazonaws.com' (ED25519) to the
Welcome to Ubuntu 24.04 LTS (GNU/Linux 6.8.0-1008-aws x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/pro

 System information as of Tue Jun  4 19:49:02 UTC 2024

  System load:  0.0               Temperature:           -273.1 C
  Usage of /:   5.5% of 28.02GB   Processes:             155
  Memory usage: 0%                Users logged in:       0
  Swap usage:   0%                IPv4 address for ens5: 172.31.29.105

Expanded Security Maintenance for Applications is not enabled.

0 updates can be applied immediately.

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status


The list of available updates is more than a week old.
To check for new updates run: sudo apt update


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.

To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

ubuntu@ip-172-31-29-105:~$ # Download
curl -O https://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/current/sratoolkit.current-ubuntu64.tar.gz

# Unzip and Untar
tar -zxvf sratoolkit.current-ubuntu64.tar
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 89.1M  100 89.1M    0     0  34.4M      0  0:00:02  0:00:02 --:--:-- 34.5M
```

```
tar (child): sratoolkit.current-ubuntu64.tar: Cannot open: No such file or directory
tar (child): Error is not recoverable: exiting now
tar: Child returned status 2
tar: Error is not recoverable: exiting now
ubuntu@ip-172-31-29-105:~$ ls
sratoolkit.current-ubuntu64.tar.gz
ubuntu@ip-172-31-29-105:~$ tar -zxvf sratoolkit.current-ubuntu64.tar
tar (child): sratoolkit.current-ubuntu64.tar: Cannot open: No such file or directory
tar (child): Error is not recoverable: exiting now
tar: Child returned status 2
tar: Error is not recoverable: exiting now
ubuntu@ip-172-31-29-105:~$ tar -zxvf sratoolkit.current-ubuntu64.tar.gz
sratoolkit.3.1.1-ubuntu64/
sratoolkit.3.1.1-ubuntu64/README.md
sratoolkit.3.1.1-ubuntu64/README-vdb-config
sratoolkit.3.1.1-ubuntu64/schema/
sratoolkit.3.1.1-ubuntu64/schema/vdb/
sratoolkit.3.1.1-ubuntu64/schema/vdb/vdb.vschema
sratoolkit.3.1.1-ubuntu64/schema/vdb/built-in.vschema
sratoolkit.3.1.1-ubuntu64/schema/insdc/
sratoolkit.3.1.1-ubuntu64/schema/insdc/insdc.vschema
sratoolkit.3.1.1-ubuntu64/schema/insdc/sra.vschema
sratoolkit.3.1.1-ubuntu64/schema/insdc/seq.vschema
sratoolkit.3.1.1-ubuntu64/schema/sra/
sratoolkit.3.1.1-ubuntu64/schema/sra/abi.vschema
sratoolkit.3.1.1-ubuntu64/schema/sra/pevents.vschema
sratoolkit.3.1.1-ubuntu64/schema/sra/generic-fastq.vschema
sratoolkit.3.1.1-ubuntu64/schema/sra/454.vschema
sratoolkit.3.1.1-ubuntu64/schema/sra/pacbio.vschema
sratoolkit.3.1.1-ubuntu64/schema/sra/illumina.vschema
sratoolkit.3.1.1-ubuntu64/schema/sra/helicos.vschema
sratoolkit.3.1.1-ubuntu64/schema/sra/nanopore.vschema
sratoolkit.3.1.1-ubuntu64/schema/sra/ion-torrent.vschema
sratoolkit.3.1.1-ubuntu64/schema/ncbi/
sratoolkit.3.1.1-ubuntu64/schema/ncbi/clip.vschema
sratoolkit.3.1.1-ubuntu64/schema/ncbi/seq-graph.vschema
sratoolkit.3.1.1-ubuntu64/schema/ncbi/stats.vschema
sratoolkit.3.1.1-ubuntu64/schema/ncbi/ncbi.vschema
sratoolkit.3.1.1-ubuntu64/schema/ncbi/wgs-contig.vschema
sratoolkit.3.1.1-ubuntu64/schema/ncbi/sra.vschema
sratoolkit.3.1.1-ubuntu64/schema/ncbi/spotname.vschema
sratoolkit.3.1.1-ubuntu64/schema/ncbi/pnbrdb.vschema
sratoolkit.3.1.1-ubuntu64/schema/ncbi/varloc.vschema
sratoolkit.3.1.1-ubuntu64/schema/ncbi/seq.vschema
sratoolkit.3.1.1-ubuntu64/schema/ncbi/trace.vschema
sratoolkit.3.1.1-ubuntu64/schema/align/
sratoolkit.3.1.1-ubuntu64/schema/align/refseq.vschema
sratoolkit.3.1.1-ubuntu64/schema/align/pileup-stats.vschema
sratoolkit.3.1.1-ubuntu64/schema/align/mate-cache.vschema
sratoolkit.3.1.1-ubuntu64/schema/align/seq.vschema
sratoolkit.3.1.1-ubuntu64/schema/align/align.vschema
sratoolkit.3.1.1-ubuntu64/schema/align/qstat.vschema
sratoolkit.3.1.1-ubuntu64/schema/csra2/
sratoolkit.3.1.1-ubuntu64/schema/csra2/stats.vschema
sratoolkit.3.1.1-ubuntu64/schema/csra2/reference.vschema
sratoolkit.3.1.1-ubuntu64/schema/csra2/read.vschema
sratoolkit.3.1.1-ubuntu64/schema/csra2/csra2.vschema
sratoolkit.3.1.1-ubuntu64/bin/
sratoolkit.3.1.1-ubuntu64/bin/bam-load.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/sra-stat.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/bam-load.3
sratoolkit.3.1.1-ubuntu64/bin/sra-search
sratoolkit.3.1.1-ubuntu64/bin/srapath-orig.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/sra-search.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/prefetch.3
sratoolkit.3.1.1-ubuntu64/bin/sam-dump.3
sratoolkit.3.1.1-ubuntu64/bin/sff-dump.3
sratoolkit.3.1.1-ubuntu64/bin/sff-dump
sratoolkit.3.1.1-ubuntu64/bin/fastq-dump
sratoolkit.3.1.1-ubuntu64/bin/fastq-load.3
sratoolkit.3.1.1-ubuntu64/bin/srf-load.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/sra-sort.3
sratoolkit.3.1.1-ubuntu64/bin/kdbmeta.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/sratools
sratoolkit.3.1.1-ubuntu64/bin/cache-mgr.3
sratoolkit.3.1.1-ubuntu64/bin/helicos-load.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/cache-mgr
sratoolkit.3.1.1-ubuntu64/bin/srapath
sratoolkit.3.1.1-ubuntu64/bin/vdb-config.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/kar.3
sratoolkit.3.1.1-ubuntu64/bin/srf-load.3
sratoolkit.3.1.1-ubuntu64/bin/illumina-dump
sratoolkit.3.1.1-ubuntu64/bin/illumina-load.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/srapath.3
sratoolkit.3.1.1-ubuntu64/bin/sam-dump
sratoolkit.3.1.1-ubuntu64/bin/vdb-encrypt
sratoolkit.3.1.1-ubuntu64/bin/sra-pileup.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/fastq-load.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/vdb-decrypt.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/rcexplain
sratoolkit.3.1.1-ubuntu64/bin/prefetch-orig.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/vdb-decrypt
sratoolkit.3.1.1-ubuntu64/bin/cache-mgr.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/illumina-dump.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/vdb-lock.3
sratoolkit.3.1.1-ubuntu64/bin/vdb-lock
sratoolkit.3.1.1-ubuntu64/bin/align-info.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/vdb-encrypt.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/ncbi/
sratoolkit.3.1.1-ubuntu64/bin/ncbi/default.kfg
sratoolkit.3.1.1-ubuntu64/bin/ncbi/certs.kfg
sratoolkit.3.1.1-ubuntu64/bin/ncbi/vdb-copy.kfg
sratoolkit.3.1.1-ubuntu64/bin/vdb-validate
sratoolkit.3.1.1-ubuntu64/bin/rcexplain.3
sratoolkit.3.1.1-ubuntu64/bin/vdb-copy.3
sratoolkit.3.1.1-ubuntu64/bin/vdb-config
sratoolkit.3.1.1-ubuntu64/bin/fastq-load
sratoolkit.3.1.1-ubuntu64/bin/sra-pileup-orig.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/sff-load.3
sratoolkit.3.1.1-ubuntu64/bin/kar.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/illumina-load.3
sratoolkit.3.1.1-ubuntu64/bin/vdb-dump.3
sratoolkit.3.1.1-ubuntu64/bin/test-sra.3
sratoolkit.3.1.1-ubuntu64/bin/helicos-load
sratoolkit.3.1.1-ubuntu64/bin/pacbio-load
sratoolkit.3.1.1-ubuntu64/bin/helicos-load.3
sratoolkit.3.1.1-ubuntu64/bin/fastq-dump-orig.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/abi-dump
sratoolkit.3.1.1-ubuntu64/bin/test-sra
sratoolkit.3.1.1-ubuntu64/bin/vdb-validate.3
sratoolkit.3.1.1-ubuntu64/bin/abi-load.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/dump-ref-fasta.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/illumina-dump.3
sratoolkit.3.1.1-ubuntu64/bin/pacbio-load.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/bam-load
sratoolkit.3.1.1-ubuntu64/bin/sratools.3
sratoolkit.3.1.1-ubuntu64/bin/test-sra.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/kdbmeta.3
sratoolkit.3.1.1-ubuntu64/bin/sra-sort-cg
sratoolkit.3.1.1-ubuntu64/bin/vdb-validate.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/fasterq-dump.3
sratoolkit.3.1.1-ubuntu64/bin/latf-load
sratoolkit.3.1.1-ubuntu64/bin/sra-sort.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/fastq-dump.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/abi-load.3
sratoolkit.3.1.1-ubuntu64/bin/sam-dump-orig.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/srapath.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/prefetch.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/vdb-unlock
sratoolkit.3.1.1-ubuntu64/bin/illumina-load
sratoolkit.3.1.1-ubuntu64/bin/srf-load
sratoolkit.3.1.1-ubuntu64/bin/vdb-dump.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/vdb-encrypt.3
sratoolkit.3.1.1-ubuntu64/bin/sratools.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/sra-stat.3
sratoolkit.3.1.1-ubuntu64/bin/sra-pileup.3
sratoolkit.3.1.1-ubuntu64/bin/vdb-lock.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/align-info
sratoolkit.3.1.1-ubuntu64/bin/sam-dump.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/sff-load.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/sra-search.3
sratoolkit.3.1.1-ubuntu64/bin/abi-load
sratoolkit.3.1.1-ubuntu64/bin/sra-sort-cg.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/cg-load.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/vdb-copy.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/vdb-decrypt.3
sratoolkit.3.1.1-ubuntu64/bin/fastq-dump.3
sratoolkit.3.1.1-ubuntu64/bin/dump-ref-fasta.3
sratoolkit.3.1.1-ubuntu64/bin/sff-dump.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/kar
sratoolkit.3.1.1-ubuntu64/bin/dump-ref-fasta
sratoolkit.3.1.1-ubuntu64/bin/fasterq-dump.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/cg-load
sratoolkit.3.1.1-ubuntu64/bin/vdb-copy
sratoolkit.3.1.1-ubuntu64/bin/vdb-config.3
sratoolkit.3.1.1-ubuntu64/bin/sra-sort-cg.3
sratoolkit.3.1.1-ubuntu64/bin/sra-stat
sratoolkit.3.1.1-ubuntu64/bin/abi-dump.3
sratoolkit.3.1.1-ubuntu64/bin/abi-dump.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/rcexplain.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/align-info.3
sratoolkit.3.1.1-ubuntu64/bin/sra-sort
sratoolkit.3.1.1-ubuntu64/bin/vdb-unlock.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/kdbmeta
sratoolkit.3.1.1-ubuntu64/bin/vdb-dump
sratoolkit.3.1.1-ubuntu64/bin/sra-pileup
sratoolkit.3.1.1-ubuntu64/bin/latf-load.3
sratoolkit.3.1.1-ubuntu64/bin/cg-load.3
sratoolkit.3.1.1-ubuntu64/bin/latf-load.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/prefetch
sratoolkit.3.1.1-ubuntu64/bin/fasterq-dump-orig.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/fasterq-dump
sratoolkit.3.1.1-ubuntu64/bin/pacbio-load.3
sratoolkit.3.1.1-ubuntu64/bin/vdb-dump-orig.3.1.1
sratoolkit.3.1.1-ubuntu64/bin/vdb-unlock.3
sratoolkit.3.1.1-ubuntu64/bin/sff-load
sratoolkit.3.1.1-ubuntu64/example/
sratoolkit.3.1.1-ubuntu64/example/perl/
sratoolkit.3.1.1-ubuntu64/example/perl/base-stats.pl
sratoolkit.3.1.1-ubuntu64/example/perl/mismatch-stats.pl
sratoolkit.3.1.1-ubuntu64/example/perl/gene-lookup.pl
sratoolkit.3.1.1-ubuntu64/example/perl/simplefastq.pl
sratoolkit.3.1.1-ubuntu64/example/perl/splitfastq.pl
sratoolkit.3.1.1-ubuntu64/example/perl/dump-reference.pl
sratoolkit.3.1.1-ubuntu64/example/perl/quality-stats.pl
sratoolkit.3.1.1-ubuntu64/CHANGES
sratoolkit.3.1.1-ubuntu64/README-blastn
ubuntu@ip-172-31-29-105:~$ ls
sratoolkit.3.1.1-ubuntu64  sratoolkit.current-ubuntu64.tar.gz
ubuntu@ip-172-31-29-105:~$ cd sratoolkit.3.1.1-ubuntu64/bin/
ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64/bin$ pwd
/home/ubuntu/sratoolkit.3.1.1-ubuntu64/bin
ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64/bin$ ls
abi-dump                 kar.3                  sra-stat
abi-dump.3               kar.3.1.1              sra-stat.3
abi-dump.3.1.1           kdbmeta                sra-stat.3.1.1
abi-load                 kdbmeta.3              srapath
abi-load.3               kdbmeta.3.1.1          srapath-orig.3.1.1
abi-load.3.1.1           latf-load              srapath.3
align-info               latf-load.3            srapath.3.1.1
align-info.3             latf-load.3.1.1        sratools
align-info.3.1.1         ncbi                   sratools.3
bam-load                 pacbio-load            sratools.3.1.1
bam-load.3               pacbio-load.3          srf-load
bam-load.3.1.1           pacbio-load.3.1.1      srf-load.3
cache-mgr                prefetch               srf-load.3.1.1
cache-mgr.3              prefetch-orig.3.1.1    test-sra
cache-mgr.3.1.1          prefetch.3             test-sra.3
cg-load                  prefetch.3.1.1         test-sra.3.1.1
cg-load.3                rcexplain              vdb-config
cg-load.3.1.1            rcexplain.3            vdb-config.3
dump-ref-fasta           rcexplain.3.1.1        vdb-config.3.1.1
dump-ref-fasta.3         sam-dump               vdb-copy
dump-ref-fasta.3.1.1     sam-dump-orig.3.1.1    vdb-copy.3
fasterq-dump             sam-dump.3             vdb-copy.3.1.1
fasterq-dump-orig.3.1.1  sam-dump.3.1.1         vdb-decrypt
fasterq-dump.3           sff-dump               vdb-decrypt.3
fasterq-dump.3.1.1       sff-dump.3             vdb-decrypt.3.1.1
fastq-dump               sff-dump.3.1.1         vdb-dump
fastq-dump-orig.3.1.1    sff-load               vdb-dump-orig.3.1.1
fastq-dump.3             sff-load.3             vdb-dump.3
fastq-dump.3.1.1         sff-load.3.1.1         vdb-dump.3.1.1
fastq-load               sra-pileup             vdb-encrypt
fastq-load.3             sra-pileup-orig.3.1.1  vdb-encrypt.3
fastq-load.3.1.1         sra-pileup.3           vdb-encrypt.3.1.1
helicos-load             sra-pileup.3.1.1       vdb-lock
helicos-load.3           sra-search             vdb-lock.3
helicos-load.3.1.1       sra-search.3           vdb-lock.3.1.1
illumina-dump            sra-search.3.1.1       vdb-unlock
illumina-dump.3          sra-sort               vdb-unlock.3
```

```
illumina-dump.3.1.1      sra-sort-cg            vdb-unlock.3.1.1
illumina-load            sra-sort-cg.3          vdb-validate
illumina-load.3          sra-sort-cg.3.1.1      vdb-validate.3
illumina-load.3.1.1      sra-sort.3             vdb-validate.3.1.1
kar                      sra-sort.3.1.1
ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64/bin$ prefetch --version
Command 'prefetch' not found, but can be installed with:
sudo apt install sra-toolkit
ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64/bin$ ~/sratoolkit.3.0.1-ubuntu64/bin/prefe
-bash: /home/ubuntu/sratoolkit.3.0.1-ubuntu64/bin/prefetch: No such file or directory
ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64/bin$ cd ~/sratoolkit.3.0.1-ubuntu64/bin/pre
-bash: cd: too many arguments
ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64/bin$ --version
--version: command not found
ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64/bin$ $ ~/sratoolkit.3.0.1-ubuntu64/bin/pre

/home/ubuntu/sratoolkit.3.0.1-ubuntu64/bin/prefetch : 3.0.1
$: command not found
-bash: /home/ubuntu/sratoolkit.3.0.1-ubuntu64/bin/prefetch: No such file or directory
ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64/bin$ cd prefetch
-bash: cd: prefetch: Not a directory
ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64/bin$ export PATH=$PATH:/home/ubuntu/sratool
ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64/bin$ prefetch --version
Command 'prefetch' not found, but can be installed with:
sudo apt install sra-toolkit
ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64/bin$ sudo apt install sra-toolkit
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
E: Unable to locate package sra-toolkit
ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64/bin$ cd ..
ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64$ prefetch --version
Command 'prefetch' not found, but can be installed with:
sudo apt install sra-toolkit
ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64$ export PATH=$PATH:/home/ubuntu/sratoolkit
ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64$ chmod +x /home/ubuntu/sratoolkit.3.1.1-ubu
ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64$ cd /home/ubuntu/sratoolkit.3.1.1-ubuntu64/
./prefetch --version

./prefetch : 3.1.1

ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64/bin$ echo $PATH
/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/usr/games:/usr/local/games:/snap/
ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64/bin$ prefetch --version

prefetch : 3.1.1

ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64/bin$ fastq-dump --version
```

```
fastq-dump : 3.1.1

ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64/bin$ echo 'export PATH=$PATH:/home/ubuntu/
source ~/.bashrc
ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64/bin$ cd ..
ubuntu@ip-172-31-29-105:~/sratoolkit.3.1.1-ubuntu64$ cd ..
ubuntu@ip-172-31-29-105:~$ cd
ubuntu@ip-172-31-29-105:~$ pwd
/home/ubuntu
ubuntu@ip-172-31-29-105:~$ prefetch SRR600956
2024-06-04T20:13:17 prefetch.3.1.1: 1) Resolving 'SRR600956'...
2024-06-04T20:13:18 prefetch.3.1.1: Current preference is set to retrieve SRA Normalized Forma
2024-06-04T20:13:18 prefetch.3.1.1: 1) Downloading 'SRR600956'...
2024-06-04T20:13:18 prefetch.3.1.1:  SRA Normalized Format file is being retrieved
2024-06-04T20:13:18 prefetch.3.1.1:  Downloading via HTTPS...
2024-06-04T20:13:40 prefetch.3.1.1:  HTTPS download succeed
2024-06-04T20:13:41 prefetch.3.1.1:  'SRR600956' is valid: 604278382 bytes were streamed from
2024-06-04T20:13:41 prefetch.3.1.1: 1) 'SRR600956' was downloaded successfully
ubuntu@ip-172-31-29-105:~$ ls
SRR600956  sratoolkit.3.1.1-ubuntu64  sratoolkit.current-ubuntu64.tar.gz
ubuntu@ip-172-31-29-105:~$ fastq-dump SRR600956
Read 25849655 spots for SRR600956
Written 25849655 spots for SRR600956
ubuntu@ip-172-31-29-105:~$ head SRR600956.fastq
@SRR600956.1 HWI-EAS486_0002:3:1:1382:1342 length=38
GTGTTCCAAATGCTGCAAATGGGTGTCAATGTATGTTA
+SRR600956.1 HWI-EAS486_0002:3:1:1382:1342 length=38
D?BCCA?BDBDBACD@=??BAAC>CBBBBBCBBBD?%%
@SRR600956.2 HWI-EAS486_0002:3:1:1382:5487 length=38
GATGATAGTTTCTTTTGCCGTTAGCACAATTTTTCCAA
+SRR600956.2 HWI-EAS486_0002:3:1:1382:5487 length=38
DCEECEAECEFFDFECEEEFFFFDB?BBADEEEEE???
@SRR600956.3 HWI-EAS486_0002:3:1:1382:4694 length=38
TGTAGGCTCCACCTCTGGGGGCAGGGCACAGACAAACA
ubuntu@ip-172-31-29-105:~$ grep -c "@SRR600956" SRR600956.fastq
25849655
ubuntu@ip-172-31-29-105:~$ prefetch SRR2156848
2024-06-04T20:18:30 prefetch.3.1.1: 1) Resolving 'SRR2156848'...
2024-06-04T20:18:30 prefetch.3.1.1: Current preference is set to retrieve SRA Normalized Forma
2024-06-04T20:18:31 prefetch.3.1.1: 1) Downloading 'SRR2156848'...
2024-06-04T20:18:31 prefetch.3.1.1:  SRA Normalized Format file is being retrieved
2024-06-04T20:18:31 prefetch.3.1.1:  Downloading via HTTPS...
2024-06-04T20:18:39 prefetch.3.1.1:  HTTPS download succeed
2024-06-04T20:18:40 prefetch.3.1.1:  'SRR2156848' is valid: 227793481 bytes were streamed from
2024-06-04T20:18:40 prefetch.3.1.1: 1) 'SRR2156848' was downloaded successfully
2024-06-04T20:18:40 prefetch.3.1.1: 'SRR2156848' has 0 unresolved dependencies
ubuntu@ip-172-31-29-105:~$ fastq-dump --split-3 SRR2156848
```

```
Read 2959900 spots for SRR2156848
Written 2959900 spots for SRR2156848
ubuntu@ip-172-31-29-105:~$ ls
SRR2156848          SRR600956.fastq
SRR2156848_1.fastq  sratoolkit.3.1.1-ubuntu64
SRR2156848_2.fastq  sratoolkit.current-ubuntu64.tar.gz
SRR600956
ubuntu@ip-172-31-29-105:~$ head SRR2156848_1.fastq
@SRR2156848.1 1 length=75
CTCGATAATCCCACTGGAAGGGCCAACAAAGTGGAAAGAGACCGGCTTTCTTGTCGCCTTTTTTTTTTTTTTTTT
+SRR2156848.1 1 length=75
@AC@-C-,CEEE8FF@8,,,,,BC78C,,,,C9,,,,,,,,;,6+6@BEF,,,;,,,8BFEEFFCBCCFEC=>=>
@SRR2156848.2 2 length=75
CGCGGAGCCCGGAGTCCGGATCTCGGCACCGCGGGACTCGTCCGAGCGATCTCCCTCCGACGCGCCGGCCGCTTC
+SRR2156848.2 2 length=75
@-6--++@CF7+:+BCF@7,8@,C,6@+8C7C++++7,,,,8,+++6++9,:<,,,9B++++4+7+++8+++488
@SRR2156848.3 3 length=75
CTCCTCGATCTCTCTCCTGAGCAGTTTTATCCCTTATCGTTTCAGACTTGCTCTTGTAGTGACTCTCATGCTCCT
ubuntu@ip-172-31-29-105:~$ grep -c "^@" SRR2156848_1.fastq
3040882
ubuntu@ip-172-31-29-105:~$ grep -c "^@" SRR2156848_2.fastq
3085591
ubuntu@ip-172-31-29-105:~$ prefetch SRR2156849 SRR2156850 SRR2156851
fastq-dump --split-3 SRR2156849 SRR2156850 SRR2156851
2024-06-04T20:22:53 prefetch.3.1.1: 1) Resolving 'SRR2156849'...
2024-06-04T20:22:53 prefetch.3.1.1: Current preference is set to retrieve SRA Normalized Forma
2024-06-04T20:22:53 prefetch.3.1.1: 1) Downloading 'SRR2156849'...
2024-06-04T20:22:53 prefetch.3.1.1:  SRA Normalized Format file is being retrieved
2024-06-04T20:22:53 prefetch.3.1.1:  Downloading via HTTPS...
2024-06-04T20:23:05 prefetch.3.1.1:  HTTPS download succeed
2024-06-04T20:23:06 prefetch.3.1.1:  'SRR2156849' is valid: 226003506 bytes were streamed from
2024-06-04T20:23:06 prefetch.3.1.1: 1) 'SRR2156849' was downloaded successfully
2024-06-04T20:23:06 prefetch.3.1.1: 'SRR2156849' has 0 unresolved dependencies
2024-06-04T20:23:06 prefetch.3.1.1: 2) Resolving 'SRR2156850'...
2024-06-04T20:23:06 prefetch.3.1.1: 2) Downloading 'SRR2156850'...
2024-06-04T20:23:06 prefetch.3.1.1:  SRA Normalized Format file is being retrieved
2024-06-04T20:23:06 prefetch.3.1.1:  Downloading via HTTPS...
2024-06-04T20:23:14 prefetch.3.1.1:  HTTPS download succeed
2024-06-04T20:23:15 prefetch.3.1.1:  'SRR2156850' is valid: 201306539 bytes were streamed from
2024-06-04T20:23:15 prefetch.3.1.1: 2) 'SRR2156850' was downloaded successfully
2024-06-04T20:23:15 prefetch.3.1.1: 'SRR2156850' has 0 unresolved dependencies
2024-06-04T20:23:15 prefetch.3.1.1: 3) Resolving 'SRR2156851'...
2024-06-04T20:23:16 prefetch.3.1.1: 3) Downloading 'SRR2156851'...
2024-06-04T20:23:16 prefetch.3.1.1:  SRA Normalized Format file is being retrieved
2024-06-04T20:23:16 prefetch.3.1.1:  Downloading via HTTPS...
2024-06-04T20:23:23 prefetch.3.1.1:  HTTPS download succeed
2024-06-04T20:23:23 prefetch.3.1.1:  'SRR2156851' is valid: 179358367 bytes were streamed from
2024-06-04T20:23:23 prefetch.3.1.1: 3) 'SRR2156851' was downloaded successfully
```

```
2024-06-04T20:23:23 prefetch.3.1.1: 'SRR2156851' has 0 unresolved dependencies
Read 2985576 spots for SRR2156849
Written 2985576 spots for SRR2156849
Read 2669778 spots for SRR2156850
Written 2669778 spots for SRR2156850
Read 2369745 spots for SRR2156851
Written 2369745 spots for SRR2156851
ubuntu@ip-172-31-29-105:~$ ls *.fastq
SRR2156848_1.fastq  SRR2156849_2.fastq  SRR2156851_1.fastq
SRR2156848_2.fastq  SRR2156850_1.fastq  SRR2156851_2.fastq
SRR2156849_1.fastq  SRR2156850_2.fastq  SRR600956.fastq
ubuntu@ip-172-31-29-105:~$ wget https://github.com/pachterlab/kallisto/releases/download/v0.44
tar -zxvf kallisto_linux-v0.44.0.tar.gz
--2024-06-04 20:24:41--  https://github.com/pachterlab/kallisto/releases/download/v0.44.0/kalli
Resolving github.com (github.com)... 140.82.116.3
Connecting to github.com (github.com)|140.82.116.3|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://objects.githubusercontent.com/github-production-release-asset-2e65be/2656290
--2024-06-04 20:24:41--  https://objects.githubusercontent.com/github-production-release-asset-
Resolving objects.githubusercontent.com (objects.githubusercontent.com)... 185.199.110.133, 18
Connecting to objects.githubusercontent.com (objects.githubusercontent.com)|185.199.110.133|:4
HTTP request sent, awaiting response... 200 OK
Length: 6668961 (6.4M) [application/octet-stream]
Saving to: ‘kallisto_linux-v0.44.0.tar.gz’

kallisto_linux-v0.4 100%[===================>]   6.36M  --.-KB/s    in 0.07s

2024-06-04 20:24:42 (91.0 MB/s) - ‘kallisto_linux-v0.44.0.tar.gz’ saved [6668961/6668961]

kallisto_linux-v0.44.0/
kallisto_linux-v0.44.0/license.txt
kallisto_linux-v0.44.0/kallisto
kallisto_linux-v0.44.0/test/
kallisto_linux-v0.44.0/test/chrom.txt
kallisto_linux-v0.44.0/test/transcripts.gtf.gz
kallisto_linux-v0.44.0/test/reads_2.fastq.gz
kallisto_linux-v0.44.0/test/transcripts.fasta.gz
kallisto_linux-v0.44.0/test/README.md
kallisto_linux-v0.44.0/test/Snakefile
kallisto_linux-v0.44.0/test/reads_1.fastq.gz
kallisto_linux-v0.44.0/README.md
ubuntu@ip-172-31-29-105:~$ export PATH=$PATH:/home/ubuntu/kallisto_linux-v0.44.0
ubuntu@ip-172-31-29-105:~$ kallisto cite
When using this program in your research, please cite

  Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L.
  Near-optimal probabilistic RNA-seq quantification,
  Nature Biotechnology 34, 525-527(2016), doi:10.1038/nbt.3519
```

```
ubuntu@ip-172-31-29-105:~$ wget ftp://ftp.ensembl.org/pub/release-67/fasta/homo_sapiens/cdna/Ho
gunzip Homo_sapiens.GRCh37.67.cdna.all.fa.gz
--2024-06-04 20:25:34--  ftp://ftp.ensembl.org/pub/release-67/fasta/homo_sapiens/cdna/Homo_sapi
           => ‘Homo_sapiens.GRCh37.67.cdna.all.fa.gz’
Resolving ftp.ensembl.org (ftp.ensembl.org)... 193.62.193.169
Connecting to ftp.ensembl.org (ftp.ensembl.org)|193.62.193.169|:21... connected.
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD (1) /pub/release-67/fasta/homo_sapiens/cdna ... done.
==> SIZE Homo_sapiens.GRCh37.67.cdna.all.fa.gz ... 59979785
==> PASV ... done.    ==> RETR Homo_sapiens.GRCh37.67.cdna.all.fa.gz ... done.
Length: 59979785 (57M) (unauthoritative)

Homo_sapiens.GRCh37 100%[===================>]  57.20M  17.9MB/s    in 3.7s

2024-06-04 20:25:41 (15.5 MB/s) - ‘Homo_sapiens.GRCh37.67.cdna.all.fa.gz’ saved [59979785]

ubuntu@ip-172-31-29-105:~$ kallisto index -i hg19.ensembl Homo_sapiens.GRCh37.67.cdna.all.fa

[build] loading fasta file Homo_sapiens.GRCh37.67.cdna.all.fa
[build] k-mer length: 31
[build] warning: clipped off poly-A tail (longer than 10)
        from 1369 target sequences
[build] counting k-mers ... done.
[build] building target de Bruijn graph ...  done
[build] creating equivalence classes ...  done
[build] target de Bruijn graph has 999022 contigs and contains 100753348 k-mers

ubuntu@ip-172-31-29-105:~$ kallisto quant -i hg19.ensembl -o SRR2156848_quant SRR2156848_1.fas

kallisto quant -i hg19.ensembl -o SRR2156849_quant SRR2156849_1.fastq SRR2156849_2.fastq

kallisto quant -i hg19.ensembl -o SRR2156850_quant SRR2156850_1.fastq SRR2156850_2.fastq

kallisto quant -i hg19.ensembl -o SRR2156851_quant SRR2156851_1.fastq SRR2156851_2.fastq

[quant] fragment length distribution will be estimated from the data
[index] k-mer length: 31
[index] number of targets: 176,981
[index] number of k-mers: 100,753,348
[index] number of equivalence classes: 671,976
[quant] running in paired-end mode
[quant] will process pair 1: SRR2156848_1.fastq
                             SRR2156848_2.fastq
[quant] finding pseudoalignments for the reads ... done
[quant] processed 2,959,900 reads, 2,563,611 reads pseudoaligned
[quant] estimated average fragment length: 190.486
```

```
[   em] quantifying the abundances ... done
[   em] the Expectation-Maximization algorithm ran for 1,057 rounds


[quant] fragment length distribution will be estimated from the data
[index] k-mer length: 31
[index] number of targets: 176,981
[index] number of k-mers: 100,753,348
[index] number of equivalence classes: 671,976
[quant] running in paired-end mode
[quant] will process pair 1: SRR2156849_1.fastq
                             SRR2156849_2.fastq
[quant] finding pseudoalignments for the reads ... done
[quant] processed 2,985,576 reads, 2,600,800 reads pseudoaligned
[quant] estimated average fragment length: 188.479
[   em] quantifying the abundances ... done
[   em] the Expectation-Maximization algorithm ran for 1,046 rounds


[quant] fragment length distribution will be estimated from the data
[index] k-mer length: 31
[index] number of targets: 176,981
[index] number of k-mers: 100,753,348
[index] number of equivalence classes: 671,976
[quant] running in paired-end mode
[quant] will process pair 1: SRR2156850_1.fastq
                             SRR2156850_2.fastq
[quant] finding pseudoalignments for the reads ... done
[quant] processed 2,669,778 reads, 2,372,309 reads pseudoaligned
[quant] estimated average fragment length: 186.747
[   em] quantifying the abundances ... done
[   em] the Expectation-Maximization algorithm ran for 969 rounds


[quant] fragment length distribution will be estimated from the data
[index] k-mer length: 31
[index] number of targets: 176,981
[index] number of k-mers: 100,753,348
[index] number of equivalence classes: 671,976
[quant] running in paired-end mode
[quant] will process pair 1: SRR2156851_1.fastq
                             SRR2156851_2.fastq
[quant] finding pseudoalignments for the reads ... done
[quant] processed 2,369,745 reads, 2,111,474 reads pseudoaligned
[quant] estimated average fragment length: 191.177
[   em] quantifying the abundances ... done
[   em] the Expectation-Maximization algorithm ran for 1,058 rounds
```

```
ubuntu@ip-172-31-29-105:~$ pwd
/home/ubuntu
ubuntu@ip-172-31-29-105:~$ client_loop: send disconnect: Broken pipe
(base) edwinruiz@Edwins-MacBook-Pro Class 17 % ls
bimm143_e1ruiz.pem
(base) edwinruiz@Edwins-MacBook-Pro Class 17 % chmod 400 "bimm143_e1ruiz.pem"
(base) edwinruiz@Edwins-MacBook-Pro Class 17 % ec2-35-86-79-243.us-west-2.compute.amazonaws.co
zsh: command not found: ec2-35-86-79-243.us-west-2.compute.amazonaws.com
(base) edwinruiz@Edwins-MacBook-Pro Class 17 % ssh -i "bimm143_e1ruiz.pem" ubuntu@ec2-35-86-79-
The authenticity of host 'ec2-35-86-79-243.us-west-2.compute.amazonaws.com (35.86.79.243)' can
ED25519 key fingerprint is SHA256:Ptz4XmSuaNkt65ielCTcPlH6CyDQAYPT/t+65mRFUTc.
This host key is known by the following other names/addresses:
    ~/.ssh/known_hosts:4: ec2-34-219-168-96.us-west-2.compute.amazonaws.com
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-35-86-79-243.us-west-2.compute.amazonaws.com' (ED25519) to the
Welcome to Ubuntu 24.04 LTS (GNU/Linux 6.8.0-1008-aws x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/pro

 System information as of Wed Jun  5 18:30:03 UTC 2024

  System load:  0.02               Temperature:           -273.1 C
  Usage of /:   54.1% of 28.02GB   Processes:             165
  Memory usage: 0%                 Users logged in:       0
  Swap usage:   0%                 IPv4 address for ens5: 172.31.29.105

 * Ubuntu Pro delivers the most comprehensive open source security and
   compliance features.

   https://ubuntu.com/aws/pro

Expanded Security Maintenance for Applications is not enabled.

0 updates can be applied immediately.

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status


The list of available updates is more than a week old.
To check for new updates run: sudo apt update

Last login: Tue Jun  4 21:03:44 2024 from 137.110.45.203
ubuntu@ip-172-31-29-105:~$ l
Homo_sapiens.GRCh37.67.cdna.all.fa  SRR2156850_quant/
SRR2156848/                         SRR2156851/
```

```
SRR2156848_1.fastq                  SRR2156851_1.fastq
SRR2156848_2.fastq                  SRR2156851_2.fastq
SRR2156848_quant/                   SRR2156851_quant/
SRR2156849/                         SRR600956/
SRR2156849_1.fastq                  SRR600956.fastq
SRR2156849_2.fastq                  hg19.ensembl
SRR2156849_quant/                   kallisto_linux-v0.44.0/
SRR2156850/                         kallisto_linux-v0.44.0.tar.gz
SRR2156850_1.fastq                  sratoolkit.3.1.1-ubuntu64/
SRR2156850_2.fastq                  sratoolkit.current-ubuntu64.tar.gz
ubuntu@ip-172-31-29-105:~$ head SRR2156848_quant/abundance.tsv
target_id   length  eff_length  est_counts  tpm
ENST00000539570 744 554.514 0   0
ENST00000576455 2046    1856.51 0   0
ENST00000510508 915 725.514 0   0
ENST00000474471 1209    1019.51 0   0
ENST00000381700 354 171.026 0   0
ENST00000445946 537 348.115 0   0
ENST00000472572 1086    896.514 0   0
ENST00000420022 462 273.947 0   0
ENST00000432593 318 138.933 0   0
ubuntu@ip-172-31-29-105:~$ head SRR2156849_quant/abundance.tsv
target_id   length  eff_length  est_counts  tpm
ENST00000539570 744 556.58  0   0
ENST00000576455 2046    1858.52 0   0
ENST00000510508 915 727.521 0   0
ENST00000474471 1209    1021.52 1   0.543502
ENST00000381700 354 172.759 0   0
ENST00000445946 537 350.005 0   0
ENST00000472572 1086    898.521 0   0
ENST00000420022 462 275.962 2   4.02372
ENST00000432593 318 140.494 0   0
```

```
(base) er@er-MacBook-Pro Class 17 % scp -r -i ./bimm143_e1ruiz.pem 'ubuntu@ec2-35-86-79-243.us

run_info.json                                 100%  362     1.9KB/s   00:00
abundance.tsv                                 100% 6300KB 153.1KB/s   00:41
abundance.h5                                  100% 1933KB  96.3KB/s   00:20
run_info.json                                 100%  362     0.0KB/s   00:16
abundance.tsv                                 100% 6273KB  48.1KB/s   02:10
abundance.h5                                  100% 1932KB  48.2KB/s   00:40
run_info.json                                 100%  362     0.0KB/s   00:21
abundance.tsv                                 100% 6281KB  38.4KB/s   02:43
abundance.h5                                  100% 1920KB   4.6KB/s   07:00
run_info.json                                 100%  362     0.0KB/s   00:21
abundance.tsv                                 100% 6254KB  38.5KB/s   02:42
abundance.h5                                    0%    0     0.0KB/s - stalled
```

```
(base) er@er-MacBook-Pro Class 17 % ls
17_Lab17_Analyzing_sequencing_data_in_the_cloud_[Extra_credit].ipynb
SRR2156848_quant
SRR2156849_quant
SRR2156850_quant
SRR2156851_quant
bimm143_e1ruiz.pem
(base) er@er-MacBook-Pro Class 17 % rm SRR2156851_quant
rm: SRR2156851_quant: is a directory
(base) edwinruiz@Edwins-MacBook-Pro Class 17 % rm -r SRR2156851_quant

(base) er@er-MacBook-Pro Class 17 % ls
17_Lab17_Analyzing_sequencing_data_in_the_cloud_[Extra_credit].ipynb
SRR2156848_quant
SRR2156849_quant
SRR2156850_quant
bimm143_e1ruiz.pem
(base) er@er-MacBook-Pro Class 17 % scp -r -i ./bimm143_e1ruiz.pem ubuntu@ec2-35-86-79-243.us-w
run_info.json                                 100%  362     5.0KB/s   00:00
abundance.tsv                                 100% 6254KB   3.3MB/s   00:01
abundance.h5                                  100% 1902KB   3.0MB/s   00:00
(base) er@er-MacBook-Pro Class 17 % ls
17_Lab17_Analyzing_sequencing_data_in_the_cloud_[Extra_credit].ipynb
SRR2156848_quant
SRR2156849_quant
SRR2156850_quant
SRR2156851_quant
bimm143_e1ruiz.pem
```

[2]:
```r
BiocManager::install("tximport")
BiocManager::install("DESeq2")
BiocManager::install("rhdf5")
```

```
'getOption("repos")' replaces Bioconductor standard repositories, see
'help("repositories", package = "BiocManager")' for details.
Replacement repositories:
    CRAN: https://cran.r-project.org

Bioconductor version 3.17 (BiocManager 1.30.23), R 4.3.2 (2023-10-31)

Warning message:
"package(s) not installed when version(s) same as or greater than current; use
  `force = TRUE` to re-install: 'tximport'"
Old packages: 'backports', 'BH', 'boot', 'broom', 'bslib', 'cachem',
  'checkmate', 'cli', 'cluster', 'codetools', 'commonmark', 'cowplot', 'cpp11',
  'curl', 'data.table', 'DBI', 'deldir', 'digest', 'dotCall64', 'dqrng',
  'emmeans', 'estimability', 'fansi', 'farver', 'fastcluster', 'fastmap',
  'FNN', 'foreign', 'fs', 'future', 'future.apply', 'ggplot2', 'ggrepel',
  'ggridges', 'ggsci', 'globals', 'glue', 'gplots', 'gtable', 'hardhat',
  'hdf5r', 'highr', 'Hmisc', 'htmltools', 'htmlwidgets', 'httpuv', 'igraph',
  'ISOcodes', 'jsonlite', 'KernSmooth', 'knitr', 'later', 'lattice', 'lda',
  'listenv', 'locfit', 'markdown', 'matrixStats', 'mgcv', 'minqa', 'munsell',
  'mvtnorm', 'nlme', 'openssl', 'parallelly', 'patchwork', 'pbdZMQ', 'plotly',
  'progress', 'promises', 'quanteda', 'quantreg', 'R.oo', 'Rcpp', 'RcppAnnoy',
  'RcppArmadillo', 'RcppEigen', 'RcppHNSW', 'RCurl', 'readr', 'repr',
  'reticulate', 'rlang', 'rmarkdown', 'rpart', 'RSQLite', 'rstudioapi',
  'Rtsne', 'sass', 'Seurat', 'SeuratObject', 'shape', 'shiny', 'sp', 'SparseM',
  'spatstat.data', 'spatstat.explore', 'spatstat.geom', 'spatstat.random',
  'stm', 'stringi', 'survival', 'tidyr', 'tidyselect', 'tinytex', 'uuid',
  'uwot', 'vctrs', 'viridis', 'vroom', 'WGCNA', 'withr', 'xfun', 'xml2', 'yaml'

'getOption("repos")' replaces Bioconductor standard repositories, see
'help("repositories", package = "BiocManager")' for details.
Replacement repositories:
    CRAN: https://cran.r-project.org

Bioconductor version 3.17 (BiocManager 1.30.23), R 4.3.2 (2023-10-31)

Warning message:
"package(s) not installed when version(s) same as or greater than current; use
  `force = TRUE` to re-install: 'DESeq2'"
Old packages: 'backports', 'BH', 'boot', 'broom', 'bslib', 'cachem',
  'checkmate', 'cli', 'cluster', 'codetools', 'commonmark', 'cowplot', 'cpp11',
  'curl', 'data.table', 'DBI', 'deldir', 'digest', 'dotCall64', 'dqrng',
  'emmeans', 'estimability', 'fansi', 'farver', 'fastcluster', 'fastmap',
  'FNN', 'foreign', 'fs', 'future', 'future.apply', 'ggplot2', 'ggrepel',
  'ggridges', 'ggsci', 'globals', 'glue', 'gplots', 'gtable', 'hardhat',
  'hdf5r', 'highr', 'Hmisc', 'htmltools', 'htmlwidgets', 'httpuv', 'igraph',
  'ISOcodes', 'jsonlite', 'KernSmooth', 'knitr', 'later', 'lattice', 'lda',
  'listenv', 'locfit', 'markdown', 'matrixStats', 'mgcv', 'minqa', 'munsell',
  'mvtnorm', 'nlme', 'openssl', 'parallelly', 'patchwork', 'pbdZMQ', 'plotly',
  'progress', 'promises', 'quanteda', 'quantreg', 'R.oo', 'Rcpp', 'RcppAnnoy',
  'RcppArmadillo', 'RcppEigen', 'RcppHNSW', 'RCurl', 'readr', 'repr',
  'reticulate', 'rlang', 'rmarkdown', 'rpart', 'RSQLite', 'rstudioapi',
  'Rtsne', 'sass', 'Seurat', 'SeuratObject', 'shape', 'shiny', 'sp', 'SparseM',
  'spatstat.data', 'spatstat.explore', 'spatstat.geom', 'spatstat.random',
  'stm', 'stringi', 'survival', 'tidyr', 'tidyselect', 'tinytex', 'uuid',
  'uwot', 'vctrs', 'viridis', 'vroom', 'WGCNA', 'withr', 'xfun', 'xml2', 'yaml'

'getOption("repos")' replaces Bioconductor standard repositories, see
'help("repositories", package = "BiocManager")' for details.
Replacement repositories:
    CRAN: https://cran.r-project.org

Bioconductor version 3.17 (BiocManager 1.30.23), R 4.3.2 (2023-10-31)
```

```
Installing package(s) 'rhdf5'

also installing the dependencies ‘Rhdf5lib’, ‘rhdf5filters’


Warning message:
"unable to access index for repository
https://bioconductor.org/packages/3.17/data/annotation/bin/macosx/big-sur-
arm64/contrib/4.3:
  cannot open URL
'https://bioconductor.org/packages/3.17/data/annotation/bin/macosx/big-sur-
arm64/contrib/4.3/PACKAGES'"
Warning message:
"unable to access index for repository
https://bioconductor.org/packages/3.17/data/experiment/bin/macosx/big-sur-
arm64/contrib/4.3:
  cannot open URL
'https://bioconductor.org/packages/3.17/data/experiment/bin/macosx/big-sur-
arm64/contrib/4.3/PACKAGES'"
Warning message:
"unable to access index for repository
https://bioconductor.org/packages/3.17/workflows/bin/macosx/big-sur-
arm64/contrib/4.3:
  cannot open URL
'https://bioconductor.org/packages/3.17/workflows/bin/macosx/big-sur-
arm64/contrib/4.3/PACKAGES'"


The downloaded binary packages are in
/var/folders/vw/6c5wjngs433234dthdjypz800000gn/T//Rtmpr9cv1A/downloaded_packages

Old packages: 'backports', 'BH', 'boot', 'broom', 'bslib', 'cachem',
  'checkmate', 'cli', 'cluster', 'codetools', 'commonmark', 'cowplot', 'cpp11',
  'curl', 'data.table', 'DBI', 'deldir', 'digest', 'dotCall64', 'dqrng',
  'emmeans', 'estimability', 'fansi', 'farver', 'fastcluster', 'fastmap',
  'FNN', 'foreign', 'fs', 'future', 'future.apply', 'ggplot2', 'ggrepel',
  'ggridges', 'ggsci', 'globals', 'glue', 'gplots', 'gtable', 'hardhat',
  'hdf5r', 'highr', 'Hmisc', 'htmltools', 'htmlwidgets', 'httpuv', 'igraph',
  'ISOcodes', 'jsonlite', 'KernSmooth', 'knitr', 'later', 'lattice', 'lda',
  'listenv', 'locfit', 'markdown', 'matrixStats', 'mgcv', 'minqa', 'munsell',
  'mvtnorm', 'nlme', 'openssl', 'parallelly', 'patchwork', 'pbdZMQ', 'plotly',
  'progress', 'promises', 'quanteda', 'quantreg', 'R.oo', 'Rcpp', 'RcppAnnoy',
  'RcppArmadillo', 'RcppEigen', 'RcppHNSW', 'RCurl', 'readr', 'repr',
  'reticulate', 'rlang', 'rmarkdown', 'rpart', 'RSQLite', 'rstudioapi',
  'Rtsne', 'sass', 'Seurat', 'SeuratObject', 'shape', 'shiny', 'sp', 'SparseM',
  'spatstat.data', 'spatstat.explore', 'spatstat.geom', 'spatstat.random',
  'stm', 'stringi', 'survival', 'tidyr', 'tidyselect', 'tinytex', 'uuid',
  'uwot', 'vctrs', 'viridis', 'vroom', 'WGCNA', 'withr', 'xfun', 'xml2', 'yaml'
```

3 Checking if any file is corrupted due to some errors i was seeing

[11]:

```r
library(rhdf5)
test_file <- "/Users/edwinruiz/ComputerScience/BIMM143/BIMM143/BIMM143/Class_17/
  ↪SRR2156851_quant/abundance.h5"
h5ls(test_file)
```

A data.frame: 14 × 5

	group	name	otype	dclass	dim
	<chr>	<chr>	<chr>	<chr>	<chr>
0	/	aux	H5I_GROUP		
1	/aux	bias_normalized	H5I_DATASET	FLOAT	4096
2	/aux	bias_observed	H5I_DATASET	INTEGER	4096
3	/aux	call	H5I_DATASET	STRING	1
4	/aux	eff_lengths	H5I_DATASET	FLOAT	176981
5	/aux	fld	H5I_DATASET	INTEGER	1000
6	/aux	ids	H5I_DATASET	STRING	176981
7	/aux	index_version	H5I_DATASET	INTEGER	1
8	/aux	kallisto_version	H5I_DATASET	STRING	1
9	/aux	lengths	H5I_DATASET	INTEGER	176981
10	/aux	num_bootstrap	H5I_DATASET	INTEGER	1
11	/aux	num_processed	H5I_DATASET	INTEGER	1
12	/aux	start_time	H5I_DATASET	STRING	1
13	/	est_counts	H5I_DATASET	FLOAT	176981

4 Downstream analysis

[3]:

```r
library(tximport)
library(DESeq2)
library(rhdf5)
library(ggplot2)
library(ggrepel)

setwd("/Users/edwinruiz/ComputerScience/BIMM143/BIMM143/BIMM143/Class_17")

folders <- dir(pattern="SRR21568*")
samples <- sub("_quant", "", folders)
files <- file.path(folders, "abundance.h5")
names(files) <- samples

txi.kallisto <- tximport(files, type = "kallisto", txOut = TRUE)
head(txi.kallisto$counts)
colSums(txi.kallisto$counts)

sum(rowSums(txi.kallisto$counts) > 0)
to.keep <- rowSums(txi.kallisto$counts) > 0
kset.nonzero <- txi.kallisto$counts[to.keep,]
keep2 <- apply(kset.nonzero, 1, sd) > 0
```

```
x <- kset.nonzero[keep2,]

#Principal Component Analysis
pca <- prcomp(t(x), scale = TRUE)
summary(pca)

plot(pca$x[,1], pca$x[,2],
     col=c("blue","blue","red","red"),
     xlab="PC1", ylab="PC2", pch=16)


pca_data <- data.frame(Sample = rownames(pca$x), PC1 = pca$x[, 1], PC2 =␣
 ↪pca$x[, 2], PC3 = pca$x[, 3])
pca_data$Condition <- c("Control", "Control", "Treatment", "Treatment")

colData <- data.frame(condition = factor(rep(c("control", "treatment"), each =␣
 ↪2)))
rownames(colData) <- colnames(txi.kallisto$counts)
y <- as.data.frame(pca$x)
y$Condition <- as.factor(colData$condition)

#ggplot for PC1 vs PC2, PC1 vs PC3 & PC2 vs PC3
ggplot(y) +
  aes(PC1, PC2, col=Condition) +
  geom_point() +
  geom_text_repel(label=rownames(y)) +
  theme_bw()

ggplot(y) +
  aes(PC1, PC3, col=Condition) +
  geom_point() +
  geom_text_repel(label=rownames(y)) +
  theme_bw()

ggplot(y) +
  aes(PC2, PC3, col=Condition) +
  geom_point() +
  geom_text_repel(label=rownames(y)) +
  theme_bw()

#OPTIONAL: Differential-expression analysis
sampleTable <- data.frame(condition = factor(rep(c("control", "treatment"),␣
 ↪each = 2)))
rownames(sampleTable) <- colnames(txi.kallisto$counts)

dds <- DESeqDataSetFromTximport(txi.kallisto, sampleTable, ~condition)
dds <- DESeq(dds)
```

```
res <- results(dds)
head(res)
```

```
1
2
3
4
```

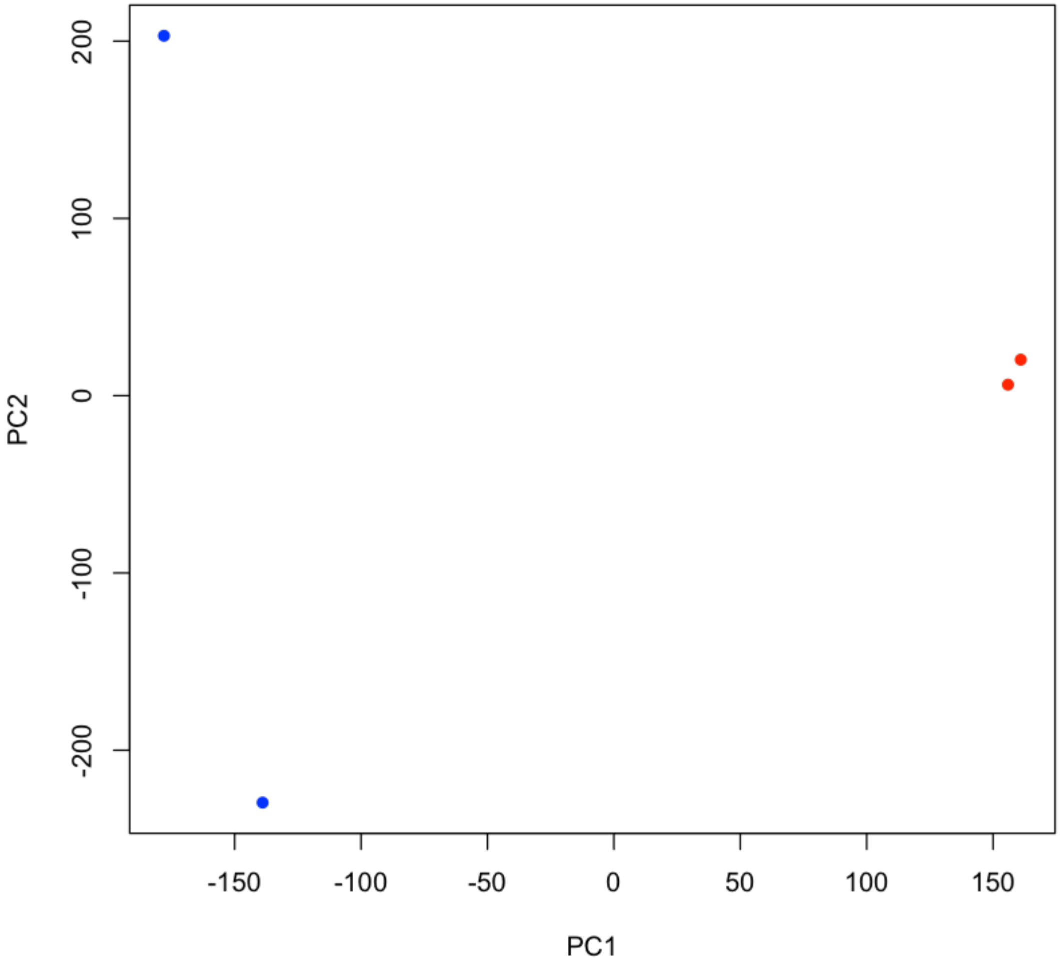
A matrix: 6 × 4 of type dbl

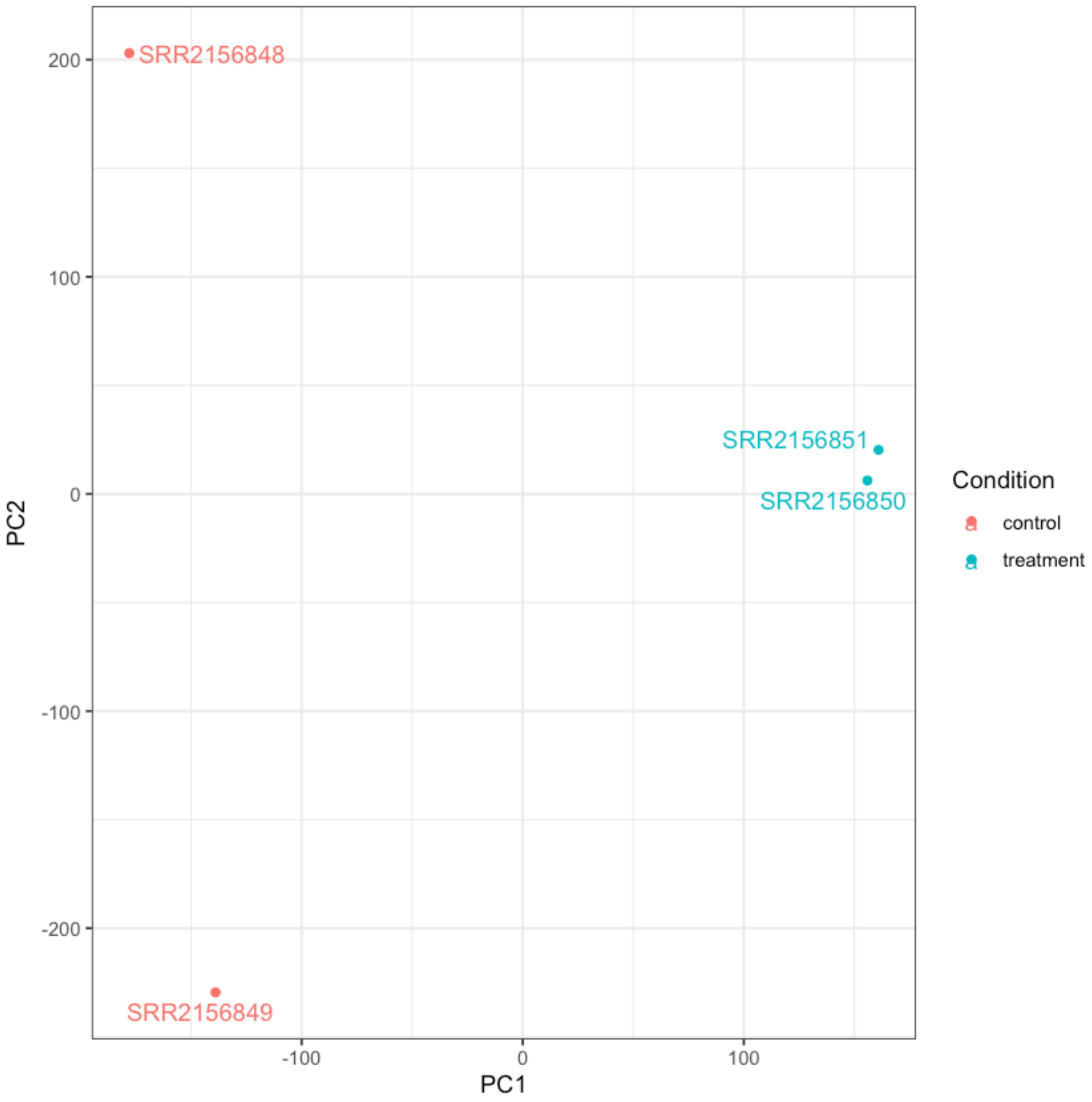
	SRR2156848	SRR2156849	SRR2156850	SRR2156851
ENST00000539570	0	0	0.00000	0
ENST00000576455	0	0	2.62037	0
ENST00000510508	0	0	0.00000	0
ENST00000474471	0	1	1.00000	0
ENST00000381700	0	0	0.00000	0
ENST00000445946	0	0	0.00000	0

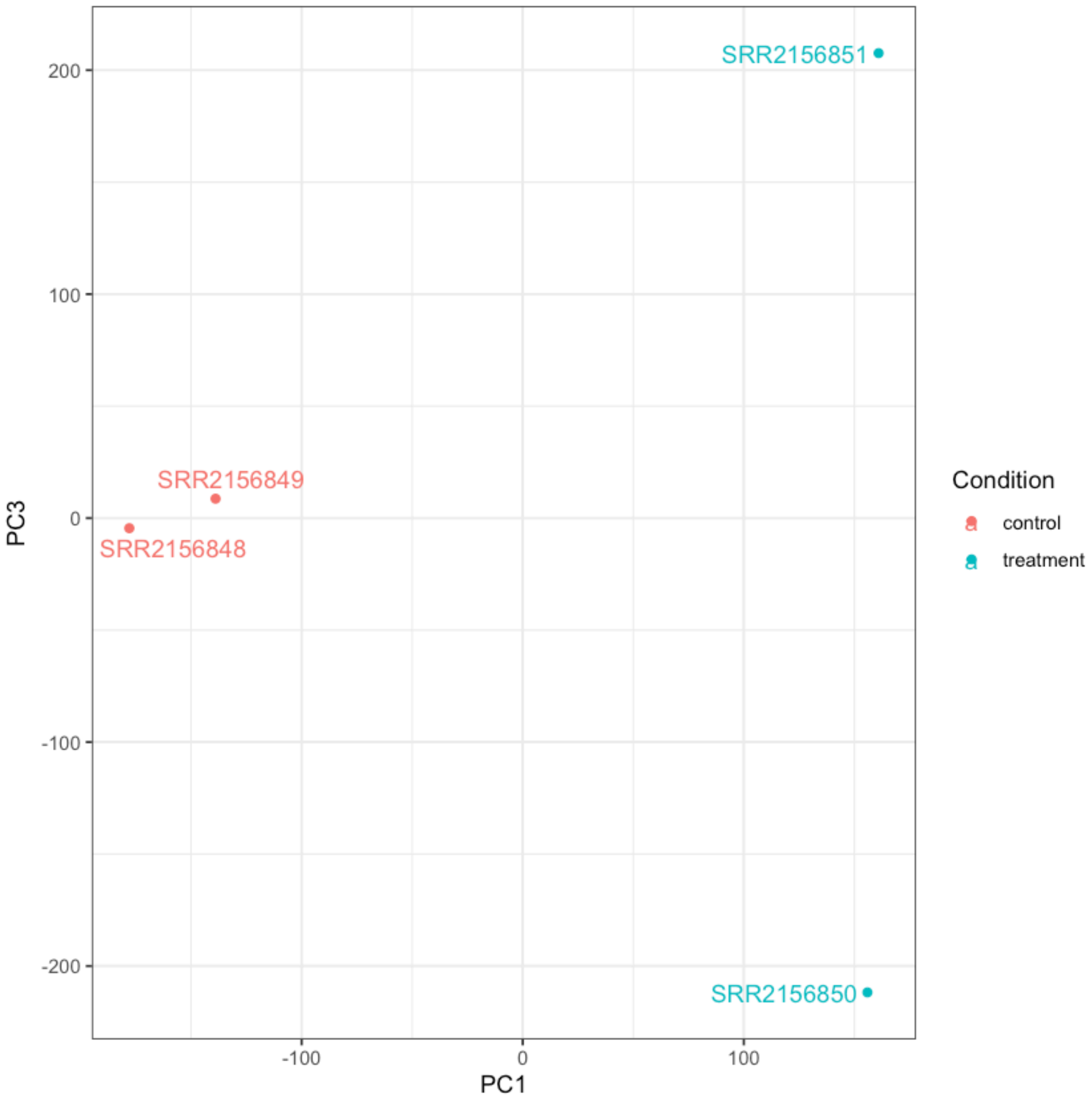
SRR2156848 2563610.99999999 **SRR2156849** 2600800 **SRR2156850** 2372309 **SRR2156851** 2111474.00000001

94561

```
Importance of components:
                            PC1      PC2      PC3   PC4
Standard deviation     183.6379 177.3605 171.3020 1e+00
Proportion of Variance   0.3568   0.3328   0.3104 1e-05
Cumulative Proportion    0.3568   0.6895   1.0000 1e+00
```

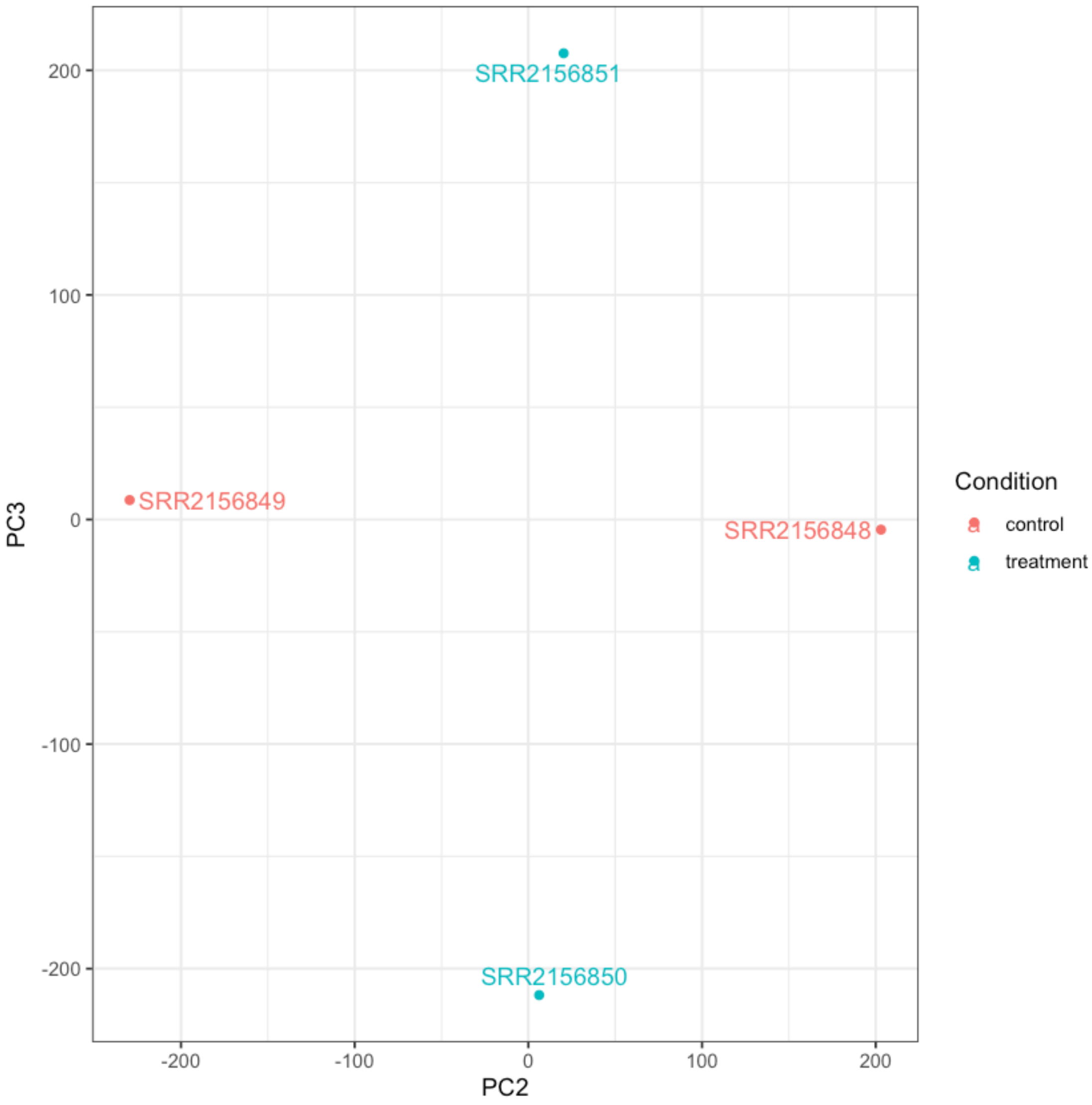

PC2
200
100
0
-100
-200
-150
-100
-50
0
50
100
150
PC1


SRR2156848
SRR2156851
SRR2156850
SRR2156849
200
100
0
-100
-200
-100
0
100
PC1
PC2
Condition
control
treatment



```
using counts and average transcript lengths from tximport

estimating size factors

using 'avgTxLength' from assays(dds), correcting for library size

estimating dispersions

gene-wise dispersion estimates

mean-dispersion relationship

-- note: fitType='parametric', but the dispersion trend was not well captured by
```

```
the
   function: y = a/x + b, and a local regression fit was automatically
substituted.
   specify fitType='local' or 'mean' to avoid this message next time.

final dispersion estimates

fitting model and testing


log2 fold change (MLE): condition treatment vs control
Wald test p-value: condition treatment vs control
DataFrame with 6 rows and 6 columns
                 baseMean log2FoldChange     lfcSE      stat    pvalue
                <numeric>      <numeric> <numeric> <numeric> <numeric>
ENST00000539570  0.000000             NA        NA        NA        NA
ENST00000576455  0.761453       3.155061   4.86052 0.6491203  0.516261
ENST00000510508  0.000000             NA        NA        NA        NA
ENST00000474471  0.484938       0.181923   4.24871 0.0428185  0.965846
ENST00000381700  0.000000             NA        NA        NA        NA
ENST00000445946  0.000000             NA        NA        NA        NA
                     padj
                <numeric>
ENST00000539570        NA
ENST00000576455        NA
ENST00000510508        NA
ENST00000474471        NA
ENST00000381700        NA
ENST00000445946        NA
```



200
100
0
-100
-200
PC3
-200
-100
0
100
200
PC2
SRR2156851
SRR2156849
SRR2156848
SRR2156850
Condition
control
treatment

[ ]: