Session 3, Lecture 1 : Exercise

- **Use Tophat to align two samples of maize short-read data**
- **Use Cufflinks to detect differentially expressed genes**

The maize genome sequence is available via: http://maizesequence.org/
We will use 2 lanes of maize Mo17 data, sequenced by the lab of Tom Brutnell, Boyce Thompson Institute, and made publicly available via GEO Submission (GSE16916)
This data has been used in the following publication:
Schnable et al. (2009) *The B73 maize genome: complexity, diversity and dynamics*. Science 326(5956): 1112-1115

Step 1.

Log into the CAC Linux server.
Create a session3 directory under your home directory on the CAC server. Copy all data files of the project to your session1 directory.

```
mkdir session3
cd session3
cp /home/gfs08/jp86/ngw2010/session3/lecture1/* ./
```

After you finish these steps, make sure you see the following files by typing "ls -l" followed by Enter key. "ls -l" command would give you the size of the file (in bytes), and last time it was modified.

maize_genome.fa (fasta file of maize genome)
s_1_sequence.txt.gz (fastq file of Illumina sequencing data, sample “zero”)
s_8_sequence.txt.gz (fastq file of Illumina sequencing data, sample “one”)
run.sh (a script for submitting a job to run the analyses)

Step 2. Submit a job to run Tophat and Cufflinks

Modify the first five lines in the run.sh file.

Modify the line "#PBS -A jp86_0004". Change "jp86_0004" to your own project name. For this project, v4 cluster (16G RAM) would work. Sometimes, the queue for the v4 cluster is very long, so you might want to use v4-64g instead.

Keep only steps 1) and 2) as shown in the run.sh file (comment out the rest of the steps with a #).
Submit the job by typing this command line followed by Enter.

```
nsub run.sh
```

The job would take a few hours to finish.

Step3. Use Cufflinks to calculate expression levels and test for differential expression.

Keep parts 4) and 5) as shown in the run.sh file (comment out the other steps using a #). Then submit the job:

```
nsub run.sh
```

Examine the output file that shows the results of differential expression at the gene-level (0_1_gene_exp.diff). How many genes are differentially expressed?

```
grep 'yes' 0_1_gene_exp.diff | wc
```

This will print out the number of differential expressed genes, by counting the lines that have ‘yes’ in them. You may examine these lines to see the FPKM and fold-change.