Session 2: Recitation

1) Log into the HPCC and go to your home directory. Try creating a directory CMSE890Sec304. Create a subdirectory Session2 (no spaces).

2) Login to a development node. Try navigating to /mnt/research/CMSE-bioinformatics/week1. Search through the GSE69360.gene-locations.txt file and determine how many chromosomes are present. Then use Wikipedia to search the “cut” and “sort” to make a new file containing only the geneids in ascending order.

3) Use Wikipedia to learn about the “grep -v” option and then use it to look through the GSE69360.gene-logcpm.mat file and find all the records where there isn’t a series of zeros (0,0,0,…). Save these to a descriptively named file in your home directory under your newly created Session2 directory. Count how many records are in this new file.

4) Navigate to /mnt/research/CMSE-bioinformatics/week3/fastq. Use “more” to look through the SRR2012208_1.fastq file. Use grep and wc -l to search for all the entries that start with “@SRR2012208”. This is the number of reads in the file. Use the list command to determine the size of this file. Use the list command to determine the size of the .fastq.gz files. Why do you think most of the files are gzipped? How much total space is needed to store all the .fastq.gz files?