Session 3 Lecture 2 Exercise (ChIP-seq)

We will use the MACS program from X. Shirley Liu's lab (Zhang, Y, et. al. (2008) Genome Biol. 9: R137) to call peaks in a ChIP-seq experiment. The original data is from a chromatin immunoprecipitation of the transcription factor STAT1 in human HeLa cells after stimulation with interferon-γ (Rozowsky, J, et. al. (2009) Nat. Biotechnol. 27: 66-75). It was downloaded from the NCBI GEO database under accession id GSE12782. The control data set is input DNA from the same cell line, also after interferon-γ stimulation. For purposes of time, we will only be using one lane of sequencing from each of the data sets.

Step 1. Log into the CAC linux server. Create session3 directory under your home directory on the CAC server. Copy all data files of the project to your session3 directory.

```
mkdir session3
cd session3
cp /home/gfs08/jp86/ngw2010/session3/lecture2/* ./
```

After you finish these steps, make sure you see the following files by typing "ls –l" followed by Enter key.

GSM320736_HeLa_STAT1_eland_results_rep1_laneA_FC302MA_20080507_s_1.txt.gz
GSM320737_IFNg_HeLa_Input_eland_results_rep1_laneA_FC305JN_20080525_s_1.txt.gz
run_macs.sh

Step 2. Submit a job to run MACS.
Modify the first five lines of the run_macs.sh file.
Modify the line "#PBS –A jp86_0004". Change "jp86_0004" to your own project name. For this project, v4 cluster will work.

Submit the job by typing this command line followed by Enter.

```
nsub run_macs.sh
```

The job should take roughly twenty minutes to finish. Note for future reference that it is much quicker (roughly 10x) if you do not run the saturation diagnostics (leave out the –diag option) and/or do not generate the wiggle files of shifted tag counts (leave out the –wig option). You can monitor the progress with the "qstat" command. After the job is finished, you will see several new files and a new directory under the session3 directory: each beginning with HeLa_STAT1. Copy all of these files and folders to your local computer.

Step 3. Analyze the output files from MACS.

HeLa_STAT1_model.pdf shows the strand-sensitive read alignments from the most enriched peaks used to build the "model" in the first part of the MACS algorithm. Does it look like they expected profile from a good ChIP experiment?

Several of the output files are Excel spreadsheets.

HeLa_STAT1_peaks.xls has information about the parameters that were used in running MACS and details about each of the peaks; specifically the genomic coordinates, width/length of the peak, the position of the "summit" of the peak (offset from the start coordinate), the number of tags in the ChIP data (not the control), the significance of the peak compared to the background (-10*log10(pvalue)), fold enrichment over the background, and the FDR value.

HeLa_STAT1_negative_peaks.xls has similar information about peaks but called after switching the experimental and control data sets. This is to give an idea of the false discovery properties of the program.

By comparing HeLa_STAT1_peaks.xls and HeLa_STAT1_negative_peaks.xls, what fraction of the total number of peaks are "noise" (i.e. how many peaks are identified in the correct comparison of data sets and how many are identified in the reversed comparison)?

HeLa_STAT1_diag.xls contains the diagnostics concerning how many peaks are identified as a function of sequencing depth and fold enrichment over background.

Do peaks of modest enrichment (less than 20 fold over background) look like they've been saturated at the current depth of sequencing?
What is the lowest fold enrichment for which further sequencing likely will not reveal new peaks?

HeLa_STAT1_peaks.bed has the coordinates and significance (-10*log10(pvalue)) for each of the peaks (similar to HeLa_STAT1_peaks.xls) in a format that be viewed in the IGV tool you used for last week's exercise or uploaded to the UCSC genome browser website.
In the directoy **HeLa_STAT1_MACS_wiggle/** there are two more directories – **control/** and **treat**/. Within each of these are files from the raw read alignments from the control and treatment/experimental data sets, respectively. These can also be viewed in IGV or uploaded to the UCSC genome browser. Since these wiggle files are big they have been broken done to one per chromosome. Upload HeLa_STAT1_peaks.bed and wiggle files from control/ and treat/ directories for one chromosome. Scan along this chromosome to see if the peak calls look reasonable. By comparing to one of the gene annotation tracks (e.g. RefSeq genes) can you identify any genes with nearby STAT1 binding?