Your group assignment

The purpose of this group project is to learn, collaborate and practice presenting your research. Some of you may learn Python and NGS analysis, some may learn some biology, and some may learn how to teach others.

The set-up:
Take a look at the sets of next-generation sequencing (NGS) data available from the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/). Pose a biological question that you will be able to address with this data. Some examples:

- Are there gene expression differences in HeLa cells versus MCF-7 cells that give clues as to the origin of these cell lines?
- How well do H3K4me3 marks in one cell type predict gene expression levels in a related but different cell type?
- What genes are dysregulated in colon cancer?

Note that this is purposefully open-ended. You can get away with doing very little work, or you can really impress us by posing a truly interesting question. Part of the challenge is to think up a good question! If you are not sure where to start, take a look at the available ENCODE data and associated papers: http://www.ncbi.nlm.nih.gov/geo/browse/?view=series&tax=9606&suppl=SRA%20Study&submitter=17528&zsort=date&display=20

The procedure:

1. Download at least two SRA Study/Experiment files from GEO. They must be from an NGS experiment, but you can choose two samples from the same study, or mix and match as you see fit. Note that the links to the actual data are at the bottom of the page describing any single sample:

Supplementary file	Size	Download	File type/resource
SRP/SRP014/SRP014481		(ftp)	SRA Study
GSE39495_RAW.tar	1.6 Gb	(ftp)(http)(custom)	TAR (of BED, BIGWIG)

Raw data provided as supplementary file

Processed data provided as supplementary file

| NLM | NIH | GEO Help | Disclaimer | Sectior

2. Wget the SRA files into the appropriate place on Gordon (this should not be your home directory).
3. Using fastq-dump on the server, convert the file to a fastq file.

4. Using Bowtie or Tophat, map the reads from each of the two samples chosen to the transcriptome (for RNA-seq) or the whole genome (ChIP-seq).
5. Use the mapped data to address the biological question you posed in (A).
    - Make sure that you move mapped data off of Gordon onto your own machine as necessary to avoid doing any analysis directly on the server. If you are analyzing the mapped data on Gordon, make sure you are using Torque job scripts to do so.
    - Feel free to use any analysis packages downstream of the data that you want, including HOMER (http://biowhat.ucsd.edu/homer/), Cufflinks (http://cufflinks.cbcb.umd.edu), MACS (http://liulab.dfci.harvard.edu/MACS/), etc.
    - Note that the GEO Accession description should include the analysis methods used by the original posters of the data.
6. Prepare a presentation (30 minutes) for Monday to present your findings to the class.

During the presentation, ensure that:

- Everyone has a chance to speak.
- The biological question is presented.
- The computational process you went through is presented. (For example, you may have generated a table of gene expression values from the mapped reads, with tag counts per RefSeq gene, normalized to RPKM. Explain how and why you did that.)
- You explain whether or not you successfully answered the question you posed.
- You explain any future studies that will be necessary to explore your question further.

Post-processing:

Each group should come up with a clever acronym to name their project, and include it in the slides.

Each group should email us a copy of the slides (PDF), along with the names of the group members. (These slides will be viewed by PIs without your narration, so try to be clear in pictures alone without being verbose...a challenge, to be sure.)

Any commentary about how the group-work went and whether all contributed equally can be sent via email to us separately.