Bootcamp Homework 2:

Goal: Align and analyze RNA-seq data

Use the same guidelines as in HW1. Make an effort to use classes and clean code in python. You will be using this code in the project!

We are going to use bowtie to align data from two replicates to the human transcriptome:

This code aligns the two files with no mismatches and takes about 5 minutes to run:
bowtie /home/jdavistu/indices/hg19_transcriptome /home/jdavistu/HW2_data/Rep1.fastq -S --al Rep1.aligned.sam -n 0

See the python file **/home/jdavistu/HW2_data/myHW2.py** for a way to run the commands from python and compute the total run time.

a) Read the bowtie manual: http://bowtie-bio.sourceforge.net/manual.shtml
b) Experiment with one of these bowtie settings (or some other that you think is relevant):
    a. -n 1 --best
    b. -n 2
c) **Perform the following analysis twice:** (once with your .sam files, and again with the files HW2_data/Rep1.aligned.sam & HW2_data/Rep1.aligned.sam)

    a. How long did it take to run each version of bowtie?
    b. How many reads aligned?
    c. For each replicate, compute RPKM for each gene *To compute the length of each transcript, you can use the hg18_refseq.fastq, which is what the bowtie index was build from.*
    d. Report the RPKM of the top 10 transcripts in the dataset (averaged over replicates).
    e. Using your discretion, compute how 'similar' the two replicates are (e.g. using correlation or coefficient of variation)
    f. Plot a histogram of the log of the RPKM (for one replicate).
    g. Plot a scatterplot of the RPKMs of one replicate versus the other.
d) Did you find a tradeoff in bowtie between speed and performance (I'm assuming that the two replicates are SUPPOSED to be highly correlated).