1. Get some general info on the yeast transcriptome.
2. Find and download the most recent Saccharomyces cerevisiae gff from Saccharomyces Genome Database
3. Create a database using gffutils (you can install gffutils via pip). Remember to only make a database once!
4. Use gffutils to figure out all the types of feature that are in the gff
5. Use gffutils to figure out which which genes have introns. What fraction of genes have introns? Print a list of gene names for genes that have introns. Protip: note that there is an intron type, and that introns have mRNAs as parents.
6. Compute the length of every gene in the yeast genome, and output a file where the first column is gene name, and the second column is length. Warning: you have to be careful for the genes with introns! You only want the sum of the CDS lengths, and you don’t want to count the intron lengths.
7. Map a yeast RNAseq experiment.
8. Download the fastq reads from https://www.dropbox.com/s/qa1uwkk8gdv5ili/SRR1177156\_1\_1000000.fastq.gzip
9. Map the reads using bowtie2 to the yeast genome, ending up with an indexed bam file
10. Did you get decent mapping statistics (e.g. what fraction of reads mapped?)
11. Quantify expression using gffutils and pysam
12. Loop over all the genes using gffutils, just like in problem 1
13. Quantify expression for each gene using pysam. You could use the pileup() method of an AlignmentFile, like you did when building the SNP caller. Alternatively, you could try to figure out the .count\_coverage() method of an AlignmentFile which could make it easier. There is documentation of count\_coverage() on the readthedocs.
14. Output the FPKM for each gene into a file where one column is a gene name, and the other column is the FPKM. Note that outputting FPKM means that you need to divide the counts of each gene by its length, and normalize by the total number of reads that map to genes (so that means you should **keep a counter of how many reads mapped to any genes** as you iterate over gene in step b).
15. Find differential expression expression between two yeast strains.
16. Download RNAseq reads from two different yeast strains, called BY and RM (you’ve actually already mapped BY in question 2, but this is slightly more data): <https://www.dropbox.com/s/17pl5zst1ypnc6t/BY.fastq.gz> and <https://www.dropbox.com/s/uz1t62saqj9mt7b/RM.fastq.gz>
17. Map the reads using bowtie2 to the yeast genome, resulting in an indexed bam file for each strain.
18. Using gffutils and pysam, test for differential expression of every gene in the genome
19. Just like in Question 3, you will need to quantify the counts of every transcript
20. You need to keep track of the quantification for both BY and RM
21. The last slide on Lecture 13 shows you how to get a p-value for just a single comparsion, so you’re going to have to figure out how to put that in a loop!
22. OUTPUT a file with mRNA name, counts in BY, counts in RM, and a p-value for if that gene is differentially expressed