DUE: Monday Mar 18 at the BEGINNING of class.
Hand In: Answer the questions on paper, number your answers, show all work (as necessary), identify key assumptions, and indicate all collaborations and assistance received or given.

Note that showing work means that if you utilize software for assistance (programs you write or stock software such as Excel or R), you should indicate as such and provide sufficient details so that I can judge the work. That may mean sending me (by email) source code or associated Excel files.

Your work must be legible -- if your handwriting isn't great, type it up and print it.

Questions

1. You and your lab mate, Eugene Yous, are performing expression-profiling experiments using RNA-Seq. You have extracted mRNA from a mouse liver. Both you and Eugene profile the same exact mRNA sample, but you decide to use polyT primer to make your cDNA whereas Eugene decides to use random priming. You obtain the exact same results across the genome except at one locus, the gene Ipt25. You find 250 reads map to Ipt25 whereas Eugene finds 45,000 reads mapping to Ipt25.

(a) Propose an explanation for the discrepancy.

(b) After scaling both data sets so that the total number of reads are identical in both yours and Eugene's experiments, what will be the effect of the difference in Ipt25 expression on the observed expression of all the OTHER genes?

2. Consider a new algorithm for predicting whether a particular RNA binding protein binds to an exon. 10,000 exons are evaluated by the prediction method and a cutoff of 2 was selected. Everything scoring above a 2 was considered positive for the RNA binding protein whereas below this level was classified as negative. These results were then compared to a gold standard method of determining whether the RNA binding protein associates with the exon. The results are shown in the following table:

	"Gold Standard" Outcome		
Prediction Method	True	False	Total
Positive	125	25	150
Negative	375	9475	9850
Total	500	9500	10,000

Calculate:

a) Sensitivity

b) Specificity

c) Positive predictive value

3. Consider an experiment where you perform RNA-seq comparing human cells grown in glucose to cells grown in galactose.

(a) Gene A changes 10-fold between these two conditions and Gene B changes 1.2-fold. Explain how it could be that the 10-fold change is statistically insignificant whereas the 1.2-fold change is statistically significant.

(b) You are looking to find regions of statistically significant differential expression, you consider two distinct ways of looking at the problem. In the first, you look at all windows of length 10 kb. In the second, you consider only the 20,000 annotated protein coding genes. Give the pros and cons of these two approaches, being sure to comment on the statistical cutoff. (Recall that the human genome is 3.0×10^9 bp.)

4. During an analysis of a promoter you identify six sites (shown below) that alter the expression of the promoter when they are deleted.

ACGGAG
ACGTGG
AGGAAG
AGGCAC
ACGCAC
AGGGAC

(a) Fill in the nucleotide count matrix, N(b,i) for this multiple alignment. (Where i is the position in the sequence and b is the identity of the nucleotide.) Since the

number of sites is small be sure to include pseudocounts summing to one in each column, distributed according to the background nucleotide frequencies in the genome. Assume you are working in a genome where %A=%T=20%, and %G=%C=30%.

Count matrix (including pseudocounts):

	1	2	3	4	5	6
A						
C						
G						
T						

(b) Now convert the above counts matrix into a probability matrix, recalling that

$$\mathrm{P(b,i)} = \mathrm{N(b,i)} / \sum_{k=1}^{4} N(k,i)$$

Probability Matrix:

	1	2	3	4	5	6
A						
C						
G						
T						

(c) Now convert the above counts matrix into a scoring matrix, recalling that S(b,i)=log[P(b,i)/P(b)].

Scoring Matrix:

	1	2	3	4	5	6
A						
C						
G						
T						

(d) Consider the following two new sequences:

Sequence 1: TCGGAG
Sequence 2: ACTGAG

Based on the scoring scheme determined in part B, which of these two sequences is a better fit to this motif model?

5. You notice that only 60% of the peaks detected by ChIP-Seq have the known transcription factor motif nearby.

(a) Give at least two explanations for the remaining 40%.

(b) How would you test (experimentally or computationally) for the explanations you suggested in part A?

6. (Advanced) Watch Lior Pachter's 2013 Keynote at Genome Informatics: https://youtu.be/5NiFibnbE8o

Entitled, "Stories from the Supplement". (Note that the sound quality is a bit poor, and the lecture is roughly 40 minutes + 5 min Q&A.)

Then answer the following questions based on Dr. Pachter's lecture:

a) According to Pachter, what are the two fundamental problems necessary to solve the inverse problem? (He says this is a chicken and egg problem.)

b) Pachter says that throwing away ambiguous data isn't a bad way to get an estimate on the expression levels of genes, but what does he state is the problem with this approach?

c) How big is the *Seq list that Pachter maintains? [Note I'm looking for the length TODAY (at the time of the talk he says it is 52), so follow the URL he gives!]

d) What does Pachter mean by "No sample is an island"? Why is this useful from a computational stand point?

e) Define impute (a word Pachter uses) and explain why he says "it will make you queasy".

f) Why is RPKM/FPKM a metric he doesn't like? (Even though he was the one who introduced FPKM!) So what metric is better?

g) What is the major complaint that Pachter has about peer review in bioinformatics?