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.

The predominant form of Ipt25 may lack a poly-A tail.

(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?

Because Ipt25 is a larger fraction of the total reads for Eugene, the apparent expression level of the other genes will go **down** relative to your samples. This is somewhat an artifact of the normalization (as everything must be a “fraction” of the total reads and Ipt25 doesn’t occupy the same fraction in the two methods).

2. Moved to homework #5 !!!

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.

If the 10 fold change reflects small read counts (for example 10 reads compared to 1) whereas the 1.2 fold change reflects substantial read depth (1200/1000 reads) then you have more confidence in the 1.2 fold change (hence it 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.)

The gene based comparison is ~20,000 comparisons. The window based comparison is ~300,000 comparisons. Using a simple bonferroni correction to your cutoff, you would have over a 10 fold smaller cutoff for the window method to obtain the same false positive rate. The windowing method will also struggle to properly detect boundaries (imagine an element that is half in one window and half in the next, would you call it since it's signal is "washed out" in both halves by the portion not changing?). However, a windowing method can detect differential transcription in unannotated regions.

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. 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	6.2	0.2	0.2	1.2	5.2	0.2
C	0.3	3.3	0.3	2.3	0.3	3.3
G	0.3	3.3	6.3	2.3	1.3	3.3
T	0.2	0.2	0.2	1.2	0.2	0.2

(b) Now convert the above counts matrix into a probability matrix, recalling that $\mathrm{P(b,i)} = \mathrm{N(b,i)} / \sum_{k=1}^{4} \boldsymbol{N(k,i)}$

Probability Matrix:

	1	2	3	4	5	6
A	6.2/7 = .8857	0.0286	0.0286	0.1714	0.7429	0.0286
C	0.0429	0.4714	0.0429	0.3286	0.0429	0.4714
G	0.0429	0.4714	0.9000	0.3286	0.1857	0.4714
T	0.0286	0.0286	0.0286	0.1714	0.0286	0.0286

(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	0.646	-0.845	-0.845	-0.067	0.570	-0.845
C	-0.845	0.196	-0.845	0.040	-0.845	0.196
G	-0.845	0.196	0.477	0.040	-0.208	0.196
T	-0.845	-0.845	-0.845	-0.067	-0.845	-0.845

(d) Consider the following two new sequences:

Sequence 1: TCGGAG
Sequence 2: ACTGAG

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

Seq #1: 0.634
Seq #2: 0.803

The second sequence better fits the model: notice that in the two positions that vary, sequence #2 scores better (0.646, -.845) than the two positions that vary in the sequence #1 (-.845,0.447).

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%.
Possible reasons include:

a. The known motif is substantially wrong.
b. The cutoff used to decide, "have the known motif" was too stringent.
c. Your antibody is non-specific and your ChIP-seq data is very noisy (these are false positive peaks).
d. Crosslinking gave a site where a cofactor is bound (i.e. the protein of interest doesn't actually touch DNA here but is brought to this location by a co-factor).

There are other valid answers.

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

a. Try to learn a motif from your data to see if you obtain the same motif as the "known". Specifically consider those sites without the known peak. (This may also identify a co-factor as in d.)
b. Use a second antibody and re-ChIP (a biological replicate).
c. Loosen your cutoff and see if you recover a substantial portion of the missing motifs.
d. Experimentally one could look for cofactors (yeast two hybrid or other method for looking for physical interactions).

The key to answering this question was to propose something that would be useful in testing your answers in part a. They didn't have to be technically correct or detailed, but conceptually I had to understand your reasoning.

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 47 minutes in length.)

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.)

Fragment assignment and Density deconvolution.

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?
You increase the variance on your estimates.

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).

There are 98 references at https://liorpachter.wordpress.com/seq/

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

That many modern experiments are actually massive compendia of experiments with connected bits of information (populations, conditions, etc). You are interested in what changes but the core is not changing and should be leveraged (i.e. not treated independently) by doing joint analysis of these large datasets (i.e. by something like non-negative matrix factorization).

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

Impute: assign (a value) to something by inference from the value of the products or processes to which it contributes.
imputation makes people uncomfortable because it is inferring information from data under some model rather than being directly determined (such as counted) from data.

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?

The proportionality constant (normalization factor) depends on the experiment. A better metric is TPM which scales the $\boldsymbol{\rho}$ by a constant factor.

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

That mathematics is relegated to the supplement and not carefully reviewed. There are useful ideas in the supplement, not just math!