DUE: Friday April 13th at the BEGINNING of class.
Hand In: Answer the questions on paper, number your answers. Your work must be legible -- if your handwriting isn't great, type it up and print it.

For full credit you must identify key assumptions and provide reasoning (or show work) behind answers. Whenever possible, partial credit will be given if adequate work is shown. Remember I encourage working together, but you MUST indicate all collaborations and/or 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), 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.

Questions 1-5 are required of all sections. Questions 6 is required of Advanced MCDB students (5520 and 7000). All questions are weighted equally in the overall grade. Those in the undergraduate section may do the advanced question for extra credit.

All problems have a maximum value of 10 points. Sub-problem values are marked when appropriate.

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) (3pt) Propose an explanation for the discrepancy.

(b) (4pt) After scaling both data sets so that the total number of reads are identical in both yours and Eugene's experiments (e.g. RPKM), what will be the effect of the difference in Ipt25 expression on the observed expression of all the OTHER genes?

(c) (3pt) Suggest an alternative normalization scheme that is more appropriate for this problem.

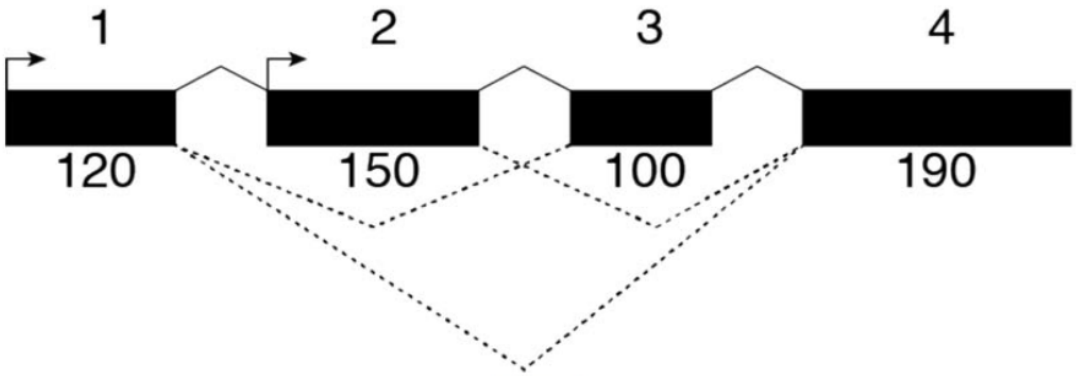
2. Consider RNA-seq experiments where you are comparing two samples.
(a) (5pt) First, you compare human cells grown in glucose to cells grown in galactose. Gene A changes 10-fold between these two conditions and Gene B changes 1.2-fold. Explain how it could be that the 10-fold change is statistically insignificant whereas the 1.2-fold change is statistically significant.

(b) (5pt) In the second, you are comparing the expression from a benign tumor to the normal adjacent cells. You have run RNA-seq in duplicate and determined that the oncogene ohNo1 is dramatically overexpressed in the tumor. However, upon closer examination, the two replicates give very different estimates for OhNo1, so you decide to do more replicates. Across many replicates it is clear that OhNo1 expression levels are

highly variable in both the tumor and normal cells. How does this knowledge influence your original conclusion that OhNo1 is overexpressed?

3. (10pts) 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 how the desired statistical cutoff is influenced by testing multiple regions. (Recall that the human genome is 3.0×10^9 bp.)

4. Consider the following gene structure:



Exons are number (1-4) and sizes are given (e.g. exon 1 is of size 120 nts). The transcript can initiate (begin) at either arrow and exons 2 and/or 3 can be spliced out (noted by dotted splice junctions).

(a) (2pt) How many possible isoforms of this gene could exist?

(b) (4pt) For each isoform, list the junction spanning RNA-seq read that would support its existence.

(c) (4pt) It was noted in class that longer read lengths will reduce the need for isoform inference algorithms. Assuming only single end reads, what is the shortest read length that would guarantee the ability to *unambiguously* identify **all** isoforms of this gene? For this question we require that a junction read must capture at least 5 bp of each adjacent exon.

5. Random short answer questions:

(a) (3pt) Explain what over dispersion is in RNA-seq data.

(b) (4pt) Explain why small indels often appear as SNP dense regions after an initial read mapping.

(c) (3pt) For each of these problems, which technique is best: nascent transcription or RNA-seq and Why?

i. Identifying the immediate transcriptional targets of a perturbation
ii. Identifying isoforms utilized.
iii. Detecting alternative 3' end (cleavage site) usage.

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.)

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).

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?