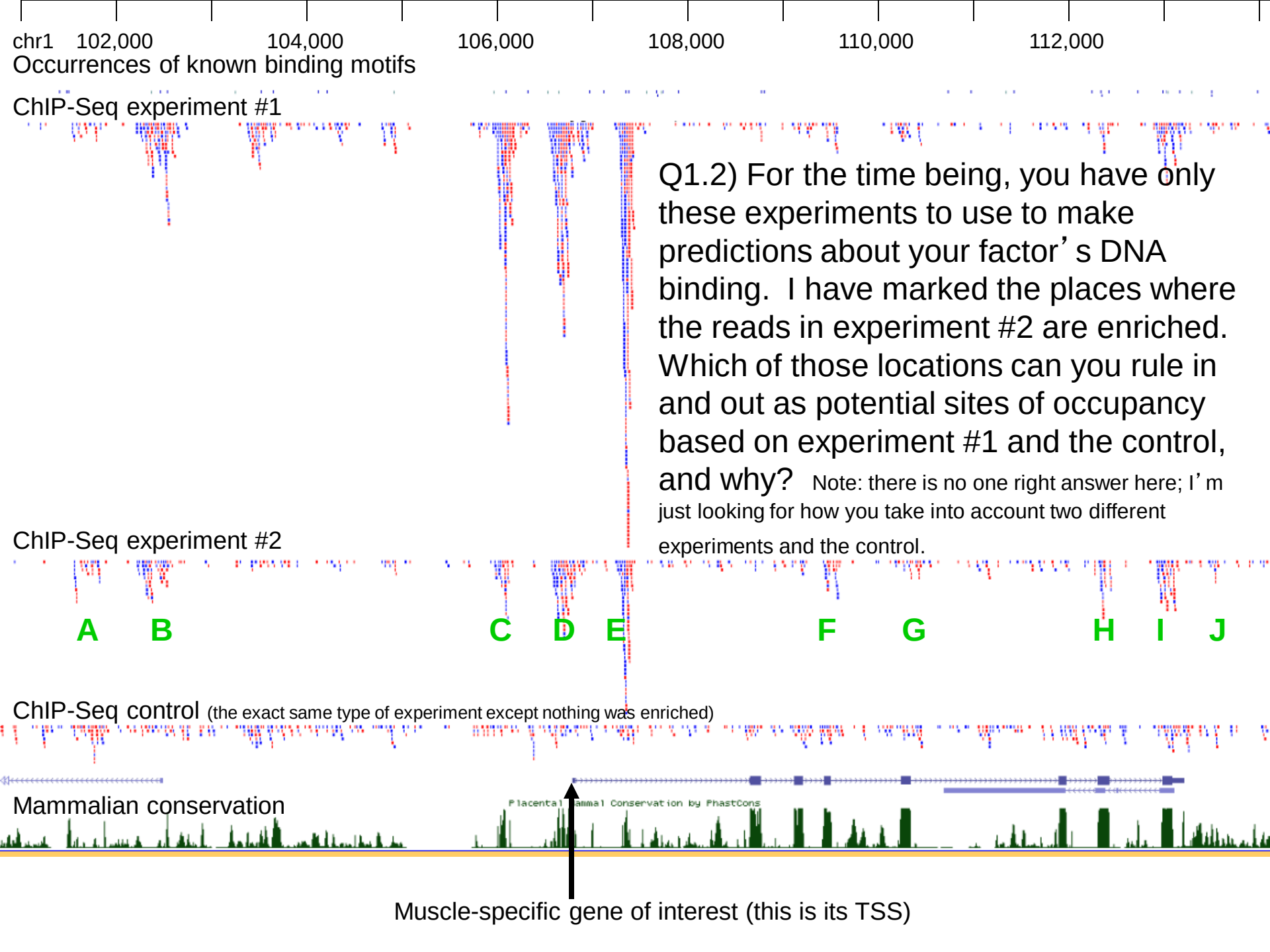
Problem 1

Question 1 intro) You are interested in a transcription factor that is known to positively regulate several muscle-specific gene during the process of muscle development. You decide to do a series of ChIP-Seq experiments in order to determine what other genes it is likely to regulate. Recall that a ChIP-Seq experiment uses an antibody specific to a factor of interest, preferentially enriches the pieces of DNA with which that factor is associated, and sequences the resulting population of DNA. After "mapping" the sequences back to a reference genome (i.e. matching them to determine where they fall on an assembled genome) , the number of reads at any point on the genome can be thought of as a probability statement that that piece of DNA was recovered in your assay (and the likelihood of the factor either directly or indirectly occupying that piece of DNA can be inferred by comparing it with read levels over the same sequence region in a control sample).

Here is a picture of what two ChIP-Seq replicate experiments and one control experiment near a gene expressed in muscle. The red and blue dashes in the experiments and the control are short (25bp) reads that were sequenced and then mapped to the reference genome. Blue reads match the plus strand of the reference genome and red reads match the minus strand of the reference genome, so you can see the orientation of the reads relative to each other. Note: this is real data from a real set of ChIP-Seq experiments, but the genome locations are simplified and therefore false.


chr1 102,000
104,000
106,000
108,000
110,000
112,000
Occurrences of known binding motifs
ChIP-Seq experiment #1
ChIP-Seq experiment #2
ChIP-Seq control (the exact same type of experiment except nothing was enriched; no antibody was used)
Mammalian conservation
Placental Mammal Conservation by PhastCons
Muscle-specific gene of interest (this is its TSS)

Q1.1) Which of the ChIP-Seq experiments gives you more information? Give two possible explanations for the difference between the two experiments.

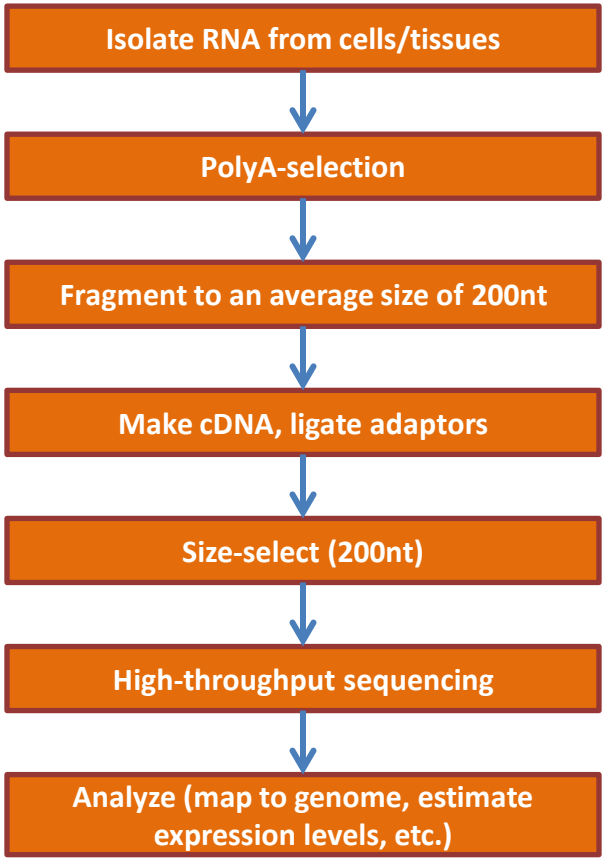

chr1 102,000
104,000
106,000
108,000
110,000
112,000
Occurrences of known binding motifs
ChIP-Seq experiment #1
Q1.2) For the time being, you have only these experiments to use to make predictions about your factor's DNA binding. I have marked the places where the reads in experiment #2 are enriched. Which of those locations can you rule in and out as potential sites of occupancy based on experiment #1 and the control, and why? Note: there is no one right answer here; I'm just looking for how you take into account two different experiments and the control.
ChIP-Seq experiment #2
A
B
C
D
E
F
G
H
I
J
ChIP-Seq control (the exact same type of experiment except nothing was enriched)
Mammalian conservation
Placental Mammal Conservation by PhastCons
Muscle-specific gene of interest (this is its TSS)

Q1.3) The top track shows the location of this factor's known DNA-binding motifs as determined by several different experimental methods. If you knew only the occurrences of these binding motifs, how well would you be able to predict the binding of this transcription factor? Give three possible reasons why you might not see factor occupancy (i.e. a "peak" of reads) every time you see a motif.

Q1.4) Based on this assay, do you think this transcription factor is likely occupying this particular gene? How would you determine this with more certainty? How would you determine if the occupancy has anything to do with regulation of the gene?

Problem 2

2.1 Below, the workflow of a typical RNA-Seq experiment is shown.



One of the primary goals of an RNA-Seq experiment is accurate quantification of gene expression levels. Note that we started with long mRNA molecules, which we fragmented and ended up with short reads mapped to the genome after sequencing. More abundant genes will have more such short sequence reads mapped to them than less abundant ones. However, longer genes will also have more mapped reads than shorter ones, if they are of the same abundance (concentration) in the sample. Therefore, an appropriate metric for measuring gene expression levels from RNA-Seq is the RPKM (Read Per Kilobase per Million mapped reads), defined as follows:

$$RPKM_i = \frac{R_i}{\left(\frac{L_i}{10^3}\right) * \frac{M}{10^6}}$$

Where R is the number of reads mapped to gene *i*, L is the length of gene *I,* and M is the total number of mapped reads

Imagine an organism that has only 20 genes. The levels of the those genes and the length of their mRNAs are shown in the table below in number of transcripts per cell in Condition A. You are also given the number of rRNA molecules in the cell. Suppose you sequenced 50 million reads from an RNA-Seq library built from RNA taken in Condition A. Assuming uniform coverage across all transcripts, expression of no alternative isoforms, 95% success in getting rid of ribosomal RNA during polyA-selection, and 36bp reads, fill up the table with the following:

2.1.1 How many reads do you expect to get from each gene?

2.1.2 What would the RPKMs be for each gene?

2.1.3 What would the average coverage of reads along the transcript be for each gene for 36bp reads?

2.1.4 What if we sequenced 75bp reads?

2.1.5 How deep would you have to sequence in order to achieve at least 5x coverage for all genes with 75bp reads?

2.1.6 What happens if the efficiency of rRNA removal becomes 80% instead of 95%?

Condition A							
#Gene	mRNA length	Molecules per cell	Number of reads	RPKM	Coverage 36bp	Coverage 75bp	RPKMs @ 80% rRNA removal efficiency
rRNA	1500	1000000					
Gene1	1000	50000					
Gene2	1000	10000					
Gene3	1000	5000					
Gene4	1000	1000					
Gene5	1000	100					
Gene6	1000	50					
Gene7	1000	10					
Gene8	1000	1					
Gene9	1000	0.1					
Gene10	10000	0.01					
Gene11	10000	50000					
Gene12	10000	10000					
Gene13	10000	5000					
Gene14	10000	1000					
Gene15	10000	100					
Gene16	10000	50					
Gene17	10000	10					
Gene18	10000	1					
Gene19	10000	0.1					
Gene20	10000	0.01					

2.1.7 Suppose we switch the organism from Condition A to Condition B, which differs from A in that the expression of Gene1 and Gene11 increases 3-fold. Assume uniform coverage across all transcripts, expression of no alternative isoforms, 95% success in getting rid of ribosomal RNA during polyA-selection, and 50 million 36bp reads and calculate the RPKMs for Condition B. Compare them to those of Condition A in the table below. Do you notice anything interesting?

	Condition B			Condition A
#Gene	mRNA length	Molecules per cell	RPKM	RPKM
rRNA	1500			
Gene1	1000			
Gene2	1000			
Gene3	1000			
Gene4	1000			
Gene5	1000			
Gene6	1000			
Gene7	1000			
Gene8	1000			
Gene9	1000			
Gene10	10000			
Gene11	10000			
Gene12	10000			
Gene13	10000			
Gene14	10000			
Gene15	10000			
Gene16	10000			
Gene17	10000			
Gene18	10000			
Gene19	10000			
Gene20	10000			

2.2 Go to the UCSC genome browser and display the titin gene (gene symbol TTN). Click on the gene and examine its characteristics (the "Sequence and Links to Tools and Databases" table would be very useful for you). Given what you learned in 2.1 about RNA-Seq experiments, can you guess why this gene may pose a challenge if one wants to very accurately quantify its expression levels though RNA-Seq. Assume that only the longest isoforms is expressed, and that the RNA was polyA-selected and fragmented the same way as in 2.1, and that all we want to know here is the expression level of the gene (i.e. we are not interested in alternative splicing, isoform reconstruction, etc.).

2.3 Given the RNA-Seq alignments and the alignment display convention shown below, can you reconstruct the most parsimonious set of transcripts (assume paired-end 75bp reads). Draw the transcript model(s) below the alignments.

pair of reads

spliced read

75bp
(shown for scaling purposes)