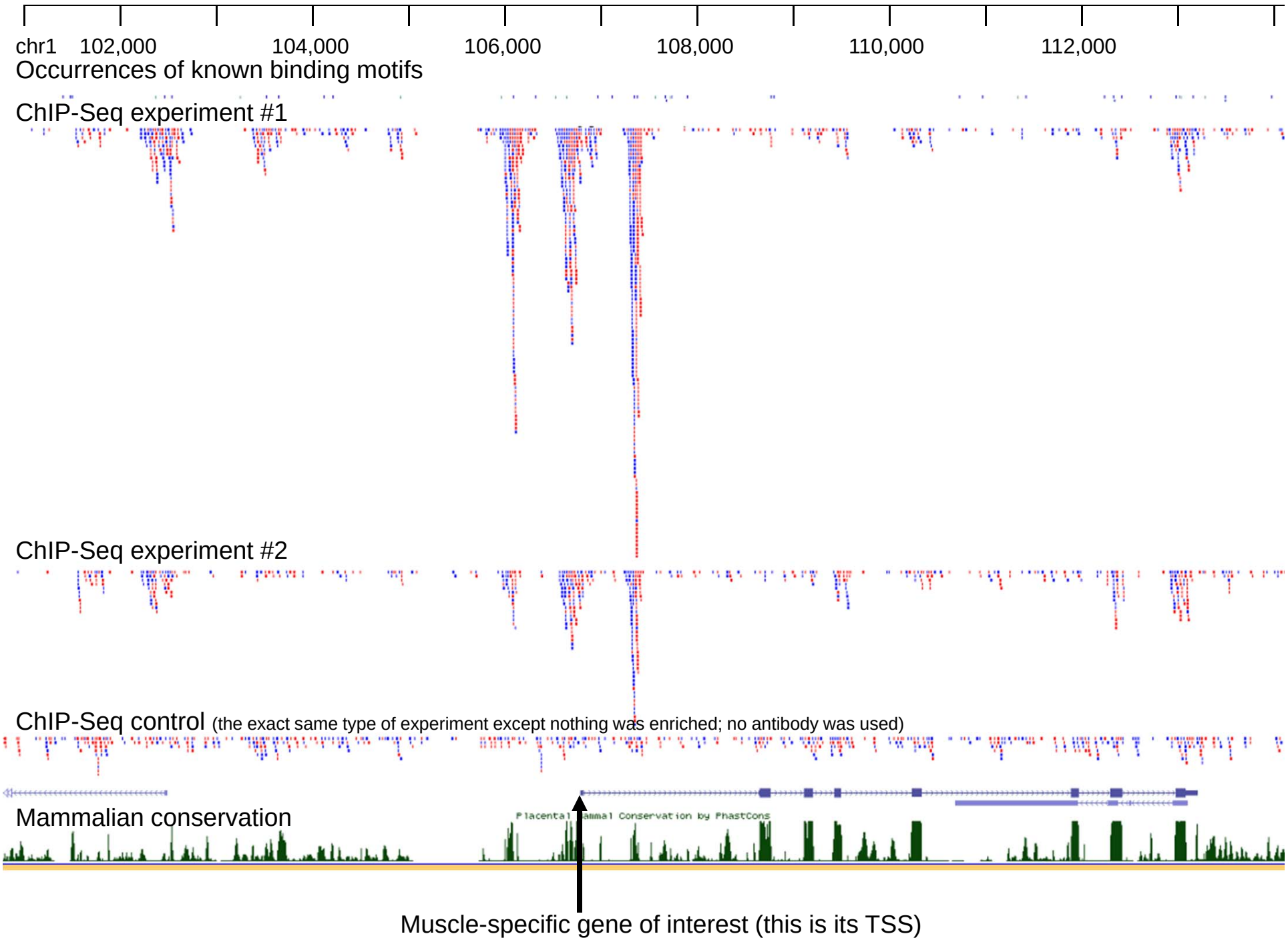
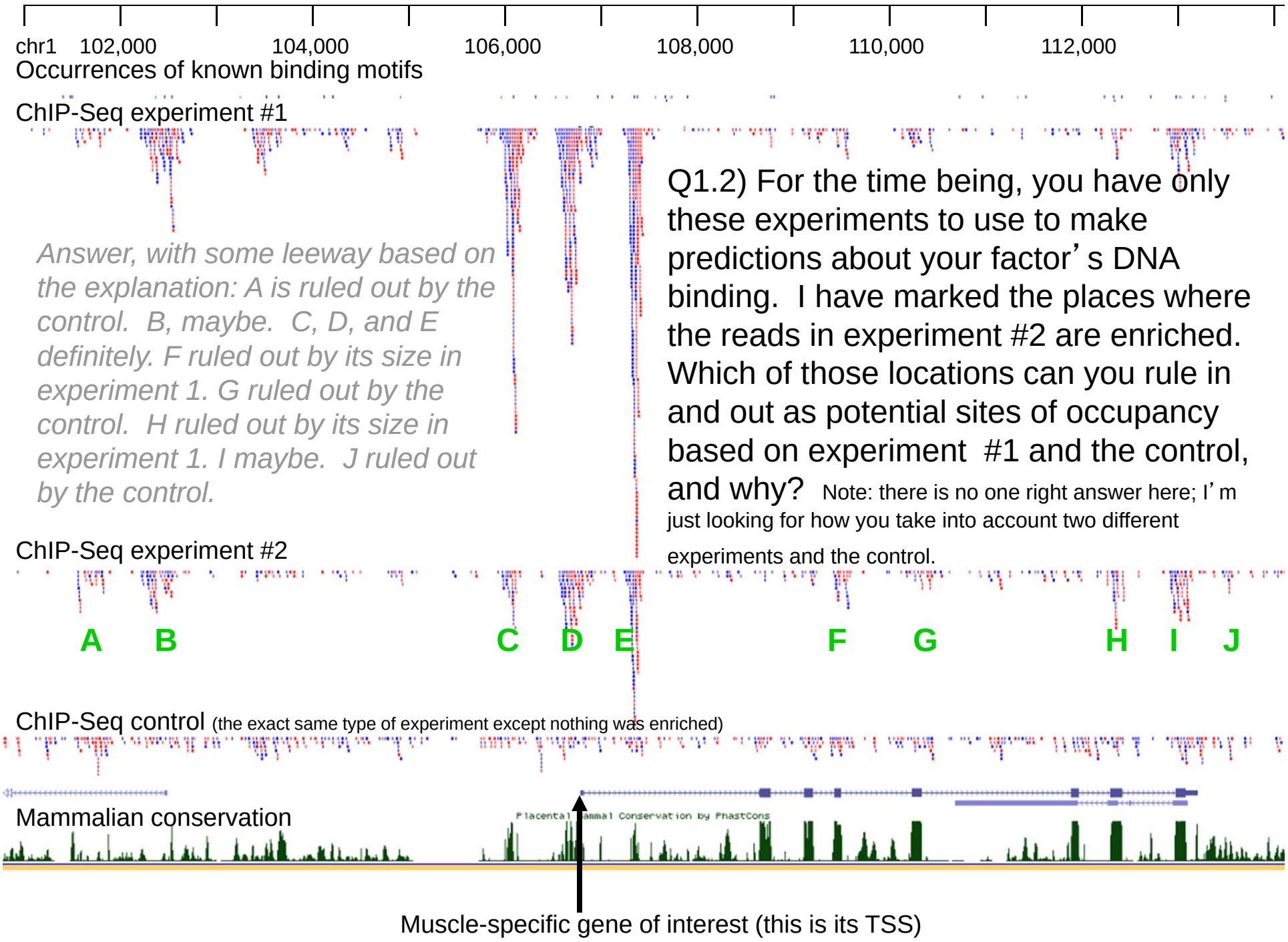
Question 1 intro) You are interested in a transcription factor that is known to positively regulate several muscle-specific gene during the process of muscle development. You decide to do a series of ChIP-Seq experiments in order to determine what other genes it is likely to regulate. Recall that a ChIP-Seq 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.

The first one gives more information. Differences include better vs. worse antibody (or older vs. newer), different cell preps, the sequencer not working as well in #2 as #1, different success during the PCR amplification step, different success at fragmenting the DNA, different amounts of cells or reagents used in the two experiments, etc.



Answer, with some leeway based on the explanation: A is ruled out by the control. B, maybe. C, D, and E definitely. F ruled out by its size in experiment 1. G ruled out by the control. H ruled out by its size in experiment 1. I maybe. J ruled out by the control.

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.

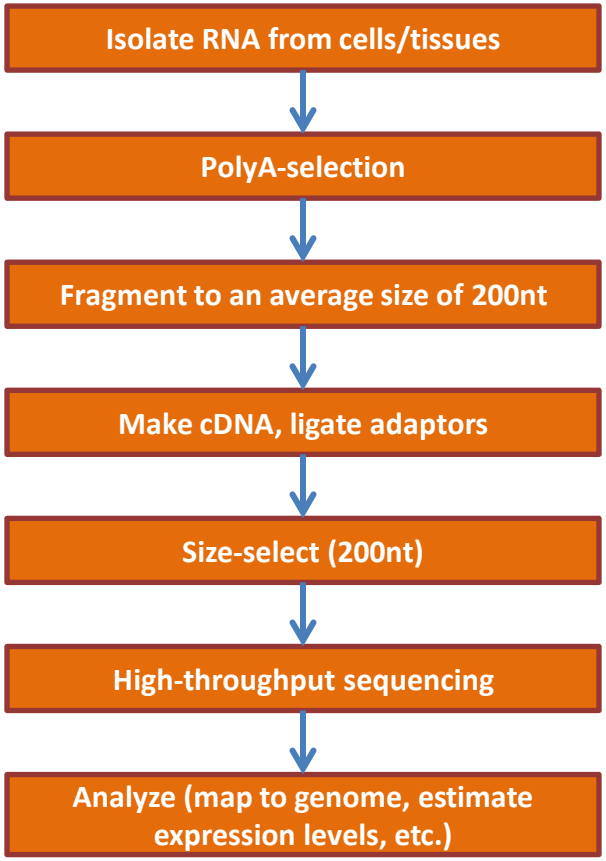
Q1.3) The top track shows the location of this factor's known DNA-binding motifs as determined by several different experimental methods. If you knew only the occurrences of these binding motifs, how well would you be able to predict the binding of this transcription factor? Give three possible reasons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

The recognition motifs for many factors are not highly predictive of factor occupancy. Possible underlying reasons are that binding requires a cofactor with its own motif; that the motifs will be occupied in some other cell type or timepoints; that the unoccupied motifs are in areas of closed or inaccessbile chromatin; or – not intended for the exercise but formally possible in real life – that the motif "known" was incorrectly determined.

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?

Probably yes, though I'd accept a "no" with a non-flawed philosophical argument. Determine it with more certainty using another experiment. For example, ChIP-Seq with a different antibody, gel shift assay, footprinting assay, etc. To determine regulatory activity, do an expression assay of some sort (BAC, plasmid, knock out the site and determine if the binding changes and/or if development messes up).

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			

Answers:

2.1.1 Given mapped read number M, gene length L_i and number of molecules per cell N_i for gene I, we expect to see the following number of reads R_i from each transcript:

$$R_i = M \frac{L_i N_i}{\sum_i^I L_i N_i}$$

2.1.2 RPKMs we calculate using the number of reads R_i obtained in 2.1.1, the number of mapped reads and the mRNA length according to the formula given before.

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

2.1.3 and 2.1.4. The expected coverage *C* is given by the following formula.

$$C_i = \frac{R_i * \text{ReadLength}}{L_i}$$

2.1.5. The sequencing depth M_{ci} at which transcript I is at coverage C can be directly derived from 2.1.1 and 2.1.3. Given that we have transcripts expressed at 0.01 copies per cell in the table, with 75bp reads one would have to sequence up to 2.97 billion reads to achieve an average of 1x coverage for those. Note that this is an extremely low concentration for the purpose of illustration of basic principle; in a homogeneous cell population the steady state level of 0.01 per cell is not physiologically meaningful for known RNA types. If only one cell in 100,000 has all the expression, that could be very meaningful to mark that cell as different from all the others.

Answers:

2.1.6 and 2.1.7 We calculate RPKMs as before. However, the RPKMs for genes other than Gene1 and Gene11, even though they are expressed at the same levels as in Condition A, have decreased in Condition B. This is because RNA-Seq provides a measure of the relative abundances of molecules in the samples, however RPKMs are directly comparable between samples only if the total number of molecules in these samples is approximately the same, which need not be true. In this case, in Condition B we have three times the number of the highest abundance genes than we do in Condition A, which leads to "compression" of the RPKMs for all other genes. The same effect is seen if significantly different amounts of rRNA are present due to varying quality of polyA-selection (or rRNA removal if one is to use one of the "ribo-minus" protocols) are present in the two samples.

Condition A							
#Gene	mRNA length	Molecules per cell	Number of reads	RPKM	Coverage 36bp	Coverage 75bp	RPKMs @ 80% rRNA removal efficiency
rRNA	1500	1000000	4671312.152	62284.162	112111.4916	233565.6076	194595.632
Gene1	1000	50000	3114208.101	62284.162	112111.4916	233565.6076	48648.9079
Gene2	1000	10000	622841.6202	12456.8324	22422.29833	46713.12152	9729.78159
Gene3	1000	5000	311420.8101	6228.4162	11211.14916	23356.56076	4864.89079
Gene4	1000	1000	62284.16202	1245.68324	2242.229833	4671.312152	972.978159
Gene5	1000	100	6228.416202	124.568324	224.2229833	467.1312152	97.2978159
Gene6	1000	50	3114.208101	62.284162	112.1114916	233.5656076	48.6489079
Gene7	1000	10	622.8416202	12.4568324	22.42229833	46.71312152	9.72978159
Gene8	1000	1	62.28416202	1.24568324	2.242229833	4.671312152	0.97297816
Gene9	1000	0.1	6.228416202	0.12456832	0.224222983	0.467131215	0.09729782
Gene10	10000	0.01	6.228416202	0.01245683	0.022422298	0.046713122	0.00972978
Gene11	10000	50000	31142081.01	62284.162	112111.4916	233565.6076	48648.9079
Gene12	10000	10000	6228416.202	12456.8324	22422.29833	46713.12152	9729.78159
Gene13	10000	5000	3114208.101	6228.4162	11211.14916	23356.56076	4864.89079
Gene14	10000	1000	622841.6202	1245.68324	2242.229833	4671.312152	972.978159
Gene15	10000	100	62284.16202	124.568324	224.2229833	467.1312152	97.2978159
Gene16	10000	50	31142.08101	62.284162	112.1114916	233.5656076	48.6489079
Gene17	10000	10	6228.416202	12.4568324	22.42229833	46.71312152	9.72978159
Gene18	10000	1	622.8416202	1.24568324	2.242229833	4.671312152	0.97297816
Gene19	10000	0.1	62.28416202	0.12456832	0.224222983	0.467131215	0.09729782
Gene20	10000	0.01	6.228416202	0.01245683	0.022422298	0.046713122	0.00972978

	Condition B			Condition A
#Gene	mRNA length	Molecules per cell	RPKM	RPKM
rRNA	1500	1000000	26277.44791	62284.162
Gene1	1000	150000	78832.34373	62284.162
Gene2	1000	10000	5255.489582	12456.8324
Gene3	1000	5000	2627.744791	6228.4162
Gene4	1000	1000	525.5489582	1245.68324
Gene5	1000	100	52.55489582	124.568324
Gene6	1000	50	26.27744791	62.284162
Gene7	1000	10	5.255489582	12.4568324
Gene8	1000	1	0.525548958	1.24568324
Gene9	1000	0.1	0.052554896	0.12456832
Gene10	10000	0.01	0.00525549	0.01245683
Gene11	10000	150000	78832.34373	62284.162
Gene12	10000	10000	5255.489582	12456.8324
Gene13	10000	5000	2627.744791	6228.4162
Gene14	10000	1000	525.5489582	1245.68324
Gene15	10000	100	52.55489582	124.568324
Gene16	10000	50	26.27744791	62.284162
Gene17	10000	10	5.255489582	12.4568324
Gene18	10000	1	0.525548958	1.24568324
Gene19	10000	0.1	0.052554896	0.12456832
Gene20	10000	0.01	0.00525549	0.01245683

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.).

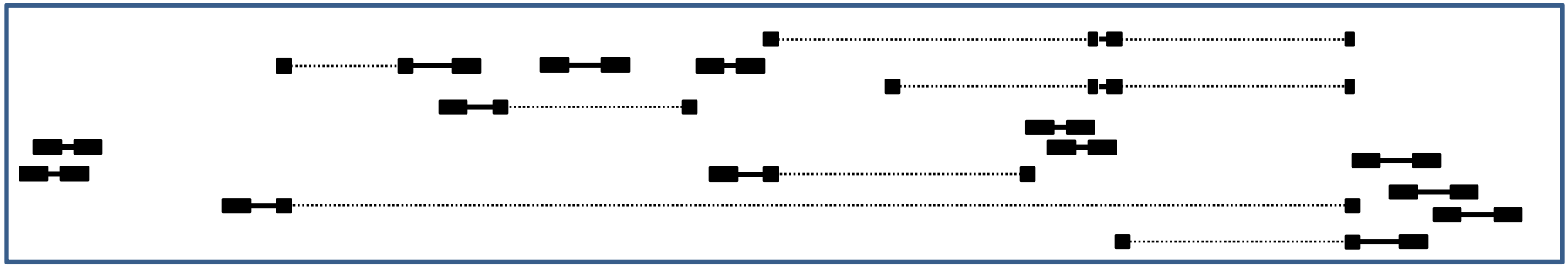
Answer: The titin gene produces an extremely long mRNA (more than 100kb long). When we polyA-select mRNA for RNA-Seq, we isolate total RNA and rely on selecting for molecules with polyA tails at their 3'-end to purify the mRNAs out of it. If during the process of RNA isolation and polyA selection, the RNA breaks, only the portion of it 3' of the break site will be represented in the library, while the portion of it 5' of the break site will be lost. If this happens, it will lead to underestimation of the levels of the gene in RPKM relative to its true expression level in the cell. Especially long genes such as titin are more likely to suffer from this effect.

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)



Answer: