Albert Liu
Bi188
Midterm

1. Genome Structure and Copy Number Variation
    a. CNVs are more likely to occur in the human genome closer to centromeres and closer to telomeres (basically the middle and ends of chromosomes).
    b. There is much greater CNV in the human genome than the mouse genome, especially with regard to neural genes? During evolution of the primate lineage you see steady increase in CNV until the orangutan? then after that CNV stays relatively stable all the way up to humans?
    c. Behavioral disorders are prominently affected by CNV? (e.g. autism?)
    d. Assuming that the top unlabeled track is that of something like H3K4Me2 (marker for enhancers) then the cis-regulatory module for the above case is most likely the broadest region of H3K4Me2 signal that also has significant conservation across species. (right half of the diagram, the second broad peak which is also the widest peak). This is because this peak looks to be ~11kB, and thus is probably the result of the duplicated cis-regulatory region of ~5.5kb that is mentioned in the problem.
    e. One alternative mechanism would be that cis-regulatory module duplication resulted in too little Bmp2 expression. This could be explained in the following manner. It is possible that due to DNA structural constraints, only the original enhancer is able to activate Bmp2 expression; in other words, the duplicated enhancer is useless because it can't properly loop around to the promoter of Bmp2. In this case, the duplicated enhancer will act in a dominant negative fashion by competing with the original enhancer for transcription factor binding, thereby reducing the likelihood that the working enhancer gets bound and activated. Consequently, expression of Bmp2 ends up being reduced.
    f. I wasn't too clear on the wording of this question:
        i. If it is asking for us to explain why healthy people generally will have duplications rather than deletions, then a potential explanation would be that there are many vital processes in the cell that require a certain threshold of gene expression in order to function properly, and thus deletions end up being deleterious/unhealthy while duplications can end up being relatively harmless (i.e. the only cost of duplication would be the cell expending a bit more resources). (e.g. tumor-suppressor genes).
        ii. If it is asking for us to explain the reverse, then a potential explanation would be that there are a lot of genes that are haplo-sufficient, i.e. the cell only needs one copy and the existence of that gene in order to survive, while over-expression stemming from duplication could significantly disrupt cell signaling processes and cause issues (e.g. oncogenes).
2. Cancer genomics
    a. The first conclusion would simply be that the COCA classification doesn't appear to be too much of a superior upgrade to the original histopathological classified groups. It would only make a significant impact for three types of tumors: BRCA tumors, as it could differentiate between C3 and C4, BLCA tumors, as it could differentiate between C1, C2,

and C8, and LUSC tumors, as it could differentiate between C1 and C2. Other than that, the histopathological and COCA classifications basically match-up one to one. On the other hand, one could make a second alternate conclusion and interpret the 1:1 alignment as a sign that COCA classification using genomic measurements is very good at identifying the type of tumor, since it matches up with the normally used histopathological identification. However, since I'm not clear just from just the figure which method is supposed to be superior, if COCA is known to be a much better identification method than histopathological classification, then the conclusion would be that COCA appears very useful for identifying rare, possibly histopathologically confusing tumors (seen by the large number of 1-5s) scattered throughout the table.

b. With regard to C2, C10, and C13, there are distinct differences in CNV throughout the chromosomes. C2 appears to have CNV throughout the entire genome. C10 appears to have very concentrated CNV, where both copies of chromosome 7 are amplified, and both copies of chromosome 10 are deleted, while the remainder of the genome is relatively untouched. Lastly, C13 basically has no CNV at all, especially when compared to all the other classifications. These 3 categories of widespread vs. concentrated vs. little/no CNV can be used to group the classifications. Under widespread CNV would be C9, C4, C2, C1, C8, and C3. Under concentrated CNV would be C7, C10, and C5. Lastly, under little/no CNV would be C6 (kind of) and C13.

c. Chromothrypsis is basically when a chromosome shatters and the DNA repair machinery freaks out and does mass re-ligation (specifically non-homologous end joining) on everything in attempt to pass the DNA-integrity checkpoint, leading to a completely rearranged chromosome. The cluster that most likely had a chromothrypsis event is C10-GBM. This is because both copies of chromosome 10 appear to be heavily deleted in almost all tumors, and this would suggest chromothrypsis since a completely rearranged chromosome would most likely have very little/no productive gene expression, and thus appear as if it was largely deleted.

d. From the genome browser we know that EGFR is on chromosome 7. From Figure 2-2 we see that chromosome 7 has significant amplification for both chromosomes in almost all glioblastomas. Thus, gene duplication is probably the most likely mutational mechanism that is contributing to mutated EGFR expression in glioblastomas, and this is also supported by the fact that when compared to other clusters C10-GBM has the highest frequency of red bands for chromosome 7, in addition to the darkest red bands (i.e. high frequency of high amplification).

3. Gene structure and expression

a. The expression data shown tells us that the splice isoform being used for CD74 is the 3rd one (from the top), with a potential for a small amount of the 1st isoform being expressed. The 1st isoform might be expressed at a very low amount because the 3rd peak is unique to the 1st isoform and its extremely small (basically it could either be noise or just a small amount). The data suggests that the 3rd isoform is being expressed because the 4th and 5th peaks are present and the 2nd isoform doesn't have them. The peak height initially suggests that there may be a 50/50 split between isoforms 2 and 3, except the 6th and 7th peaks are the same height as the 4th and 5th peaks, and the 6th and 7th peaks are shared by

all transcripts, thus suggesting that it is primarily isoform 3 and not a mixture of 2 and 3 that is being expressed. To definitively prove that it is just isoform 3 that is being produced you would want to look for reads that cross splice junctions: one that links the 2nd and 4th peak, and one that links that 4th and 5th peak. If those reads are fairly frequent then you could be more confident about the fact that it is isoform 3 that is mainly expressed.

b. The definition of RPKM is reads per kilobase of exon per million of reads. If you just calculated RPM, or reads per millions of reads, then you would incorrectly deduce that longer genes were expressed way more than shorter genes, since you aren't controlling for the length of the exon. Thus, the shortest and longest genes would be greatest affected by switching from RPKM to RPM.
c. Myog is the likely target of the transcriptional regulators shown. This is because all of the transcription factors and p300 are basically binding to what would presumably be the Myog promoter. (Marked by black arrow). Also Myog is the only gene being expressed.
d. Active transcriptional enhancers are marked with red arrows. They were chosen because they have strong transcription factor binding and they were not in protein coding sequences. Basically, there were a total of 6 peaks from transcription factor binding, one was obviously the promoter of Myog, and then two were exons of Ppfia4 (we discussed enhancers being in introns, but never exons, so I am assuming that there are no special exceptions where enhancers are in exons). As a result, 3 peaks were left and designated as enhancers, two of which had good sequence conservation but the enhancer inside an intron of Ppfia4 does not.
e. The data at this locus suggests that although sequence conservation is generally a good marker for cis-regulatory modules you must be careful because protein coding sequences will also (in some cases such as this one) have very good sequence conservation and they cannot be the site of enhancers.
f. The two main groups of categories that you see in the results are basically categories related to the neural system as well as categories related to development. This makes sense with regard to the tissue samples because the tissue samples are basically normal and diseased neural tissue (either Huntington's disease or bipolar disorder), both of which involve malfunctions in the central nervous system; Huntington's is neurodegenerative, and bipolar disorder is neurodysregulated? Thus, it makes biological sense that genes that are involved with neural function and development would be significantly over expressed and possibly be the cause of the diseased phenotypes in the analyzed neural tissue.