Bi 188 Midterm

James Ha

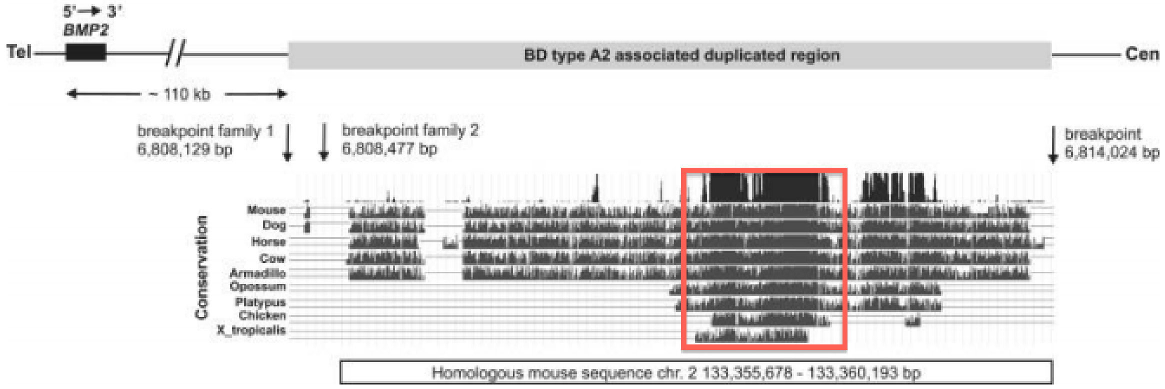
May 4, 2015

1. A. From class we saw that CNVs are more likely to occur near telomeres and centromeres.

1. B. In the mouse, there have been fewer duplication events than in humans. During primate evolution, there was a great increase in the rate of duplication events. This is clear when we compare the similarity of the duplicated regions of human, chimpanzee, and orangutan genomes. Humans are similar to chimps, but very different from orangutans, which are more distantly related!

1. C. One disease was skeletal defects due to CNVs in one of the Hedgehog genes. I believe a presentation this week also talked about CNVs of EGFR in cancer.

1. D. Sequence conservation and what I assume is a binding peak track above the conservation tracks indicate that the following region is likely the *cis*−regulatory module:



1. E. Overexpression of *Bmp*2 would likely lead to aberrant BMP2 signalling at all receptors that respond to BMP2, not just at BMPR1B. It is possible that the signal transduction cascades of these receptors and the signal transduction cascade of BMPR1B interact at some point downstream to cause the phenotype. The mutation would act in a dominant negative manner if strong inputs at the other BMP2 receptors can override the effect of normal inputs at BMPR1B (i.e., BMPR1B signalling pathway does not have overwhelmingly strong regulatory control over the signalling pathways of other BMP2 receptors).

1. F. Deletions in RefSeq genes often cause truncation of the gene products, which usually renders them non-functional. These non-functional gene products could build up and produce deleterious effects or it could be that two "doses" of the gene are required for the organism's survival/health. Either way, individuals with deletions in RefSeq genes will usually suffer from some sort of health problem. Therefore, this type of selection effect ensures that in healthy individuals we will find that deletions are less likely to associate with RefSeq genes.

2. A. It seems that many of the COCA assignments correspond almost perfectly to one of the histopathologically classified groups. For example, nearly every sample in cluster $C5$ corresponds to a sample that was classified as $KIRC$ via histopathological methods, and the same is true of $C6$ and the $UCEC$ group. So in a good number of cases, a histopathological classification of the cancer gives a pretty good indication of the genomic defects of the cancer cells.

On the other hand, there are histopathological groups that actually correspond to multiple clusters and vice versa. For example, the histopathological group $BRCA$ corresponds to two clusters $C3$ and $C4$, while the cluster $C7$ corresponds to the histopathological groups $COAD$ and $READ$. This indicates that for certain cancer types, the histopathological classification method either makes a useless distinction (when considering potential treatments) between cancer types as in the case of $COAD$ and $READ$, or it fails to make an important distinction between cancer types with distinct genomic properties as in the case of $BRCA$.

2. B. Many of the clusters have similar CNV patterns with the major difference being the relative severity of the defects in different chromosomes, but several of them have strikingly different CNV patterns. For example, $C2$ cancers have severe defects throughout the genome, whereas $C10$ cancers appear to have major CNVs at *chr* $7, 10$ with little to moderate amounts of CNVs elsewhere, and $C13$ cancers have almost no CNVs. Most of the clusters are similar to $C2$; $C9, C8, C4, C1$, and $C3$ all have a similar CNV patterns to $C2$ (especially at *chr* $7, 8$), with severe defects genome wide. $C10$, and $C13$ are fairly unique among the clusters in that $C10$ is one of the few clusters to have most of the extreme CNVs localized to a small number of chromosomes, while $C13$ is the only cluster to have no extreme CNVs anywhere.

2. C. Chromothrypsis describes a type of catastrophic DNA-damage in which one or a few chromosomes are basically shattered into many fragments. Some of the fragments are joined together in an attempt to repair the chromosome but several fragments are typically lost and some may form double-minute chromosomes. It seems most likely that cluster $C13$ suffered from chromothrypsis, since most of the genome is unharmed but there appears to be a fair amount of deletion in *chr* 7. This CNV pattern is consistent with a chromothrypsis event fragmenting *chr* 7 and causing the loss of some parts of the chromosome.

2. D. This is probably due to the duplication of *chr* 7 in $C10$ samples. No other cluster displays such a prominent amplification of *chr* 7. If the EGFR gene is on *chr* 7 (the UCSC genome browser says it is), then it would be duplicated with the chromosome. Since only two copies of EGFR are required, the other two copies are free to accumulate mutations without compromising the survival/proliferation of the cell.

3. A. The relative heights of the RNA-Seq peaks indicate that the top two splice isoforms are less common than the bottom splice isoform (the one with 8 exons). We would want to look for reads that go across splice junctions to prove that the bottom splice isoform is the most common one.

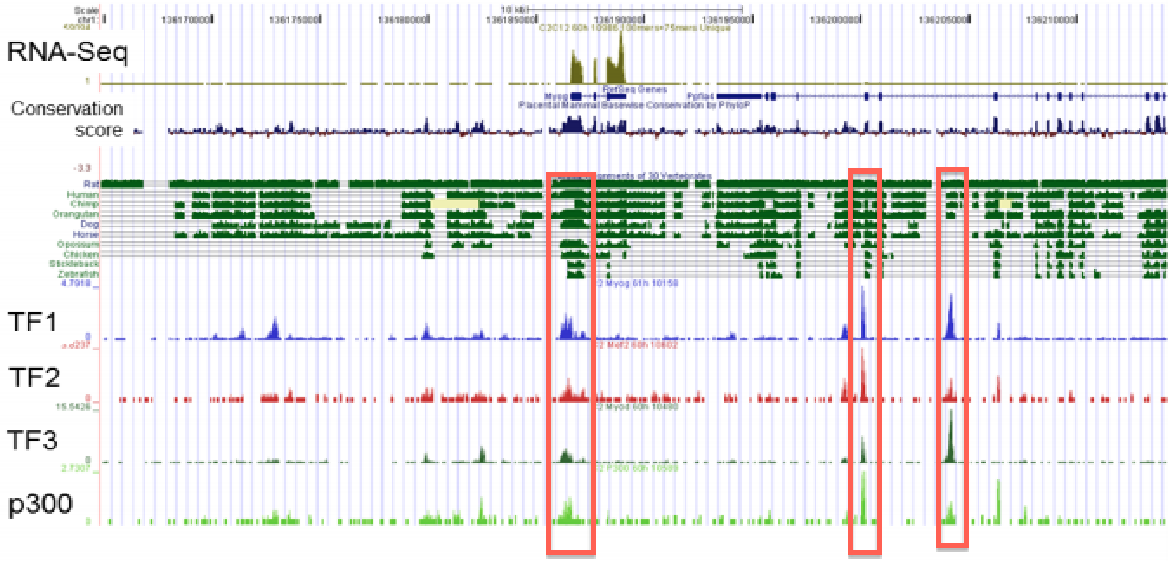
3. B. $RPKM$ is the reads per kilobase of exons per million reads. Let R be the number of reads mapped to a transcript, E be the length of all exons in the transcript in bp, and R_{tot} be the total number of reads. Then we have:

$$RPKM = \frac{R}{\frac{R_{tot}}{1000000}\frac{E}{1000}}$$

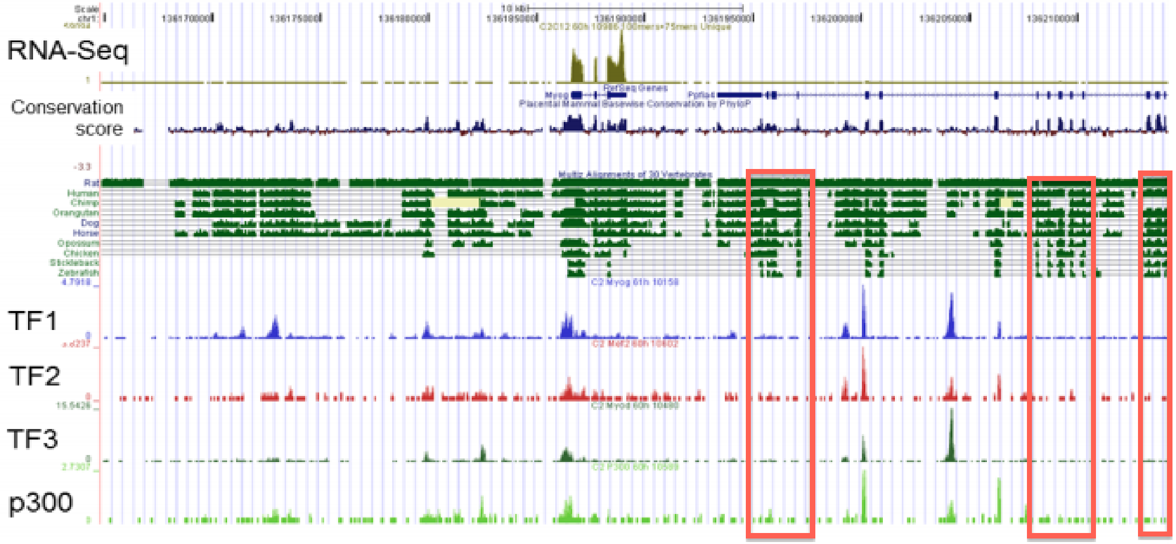
If we were to calculate expression values using RPM instead, we would be unable to compare the expression levels of genes that have very different sizes. Specifically, the largest genes may appear to have high expression values even if they are not very highly expressed, while very small genes would appear to have modest expression even if they are highly expressed!

3. C. The gene being regulated is likely the $Myog$ (Is that what that says? The letters are so small!) gene. This is because it is the only gene with RNA-Seq peaks corresponding to its exons. Also, some of the transcription factor/$p300$ binding peaks lie on the exons of $Ppfia4$ (or whatever its name is), which makes it a bit less plausible that it is the gene being transcriptionally regulated by these factors.

3. D. Based on both the high degree of sequence conservation and the strength of the transcription factor/$p300$ binding peaks, the three regions below are likely the active transcriptional enhancers.



3. E. It seems that this is not generally true since there are at least 3 groups of highly conserved sequences in the image for which there are essentially no transcription factor/$p300$ binding peaks.



3. F. Note that the tissue samples in which these genes were found to be highly expressed are all brain tissues. The functions attributed to these genes are mostly related to axon motility and neuron development. This makes a lot of biological sense, since we expect neurons to express genes related to axon motility (to form/eliminate synapses) and brain tissues are full of neurons. The neuron development functions make a bit less sense, since neurons typically do not differentiate in the adult as far as I know, but it is always possible that the tissue samples were from young brains or from cell lines. Actually, now that I think of it, these results make the most sense if the brain tissues are from young individuals.