Jessica Lam

Bi188 Midterm Examination
Spring 2015

Due **Monday, May 4rd, 12:00pm**
(as a PDF emailed to sgoh@caltech.edu, phe@caltech.edu and woldb@caltech.edu).

The exam is closed-book and closed-notes, ***though you will need internet access to use the genome browser for one question.***

You are asked to identify some specific features on a figure as well. This can be done onto a printout or you can drop the png into powerpoint or other favored program and make your annotations that way. Also, uou can draw while taking the exam and convert to electronic after time is up…..or even hand in paper.

You have 3 hours to complete the exam, though we expect that the exam can be completed well within 2 hours.

There are 3 parts to this exam that adds up to a total of 35 points.

Express your answers concisely. Most questions need only one or a few sentences.

Please read no further until you are ready to take the exam.

1. Genome structure and Copy Number Variation (total of 12 points)

A. Where are Copy Number Variations (CNVs) more likely to occur in the human genome?

CNVs are more likely to occur closer to centromeres. Certain chromosomes are also more susceptible.

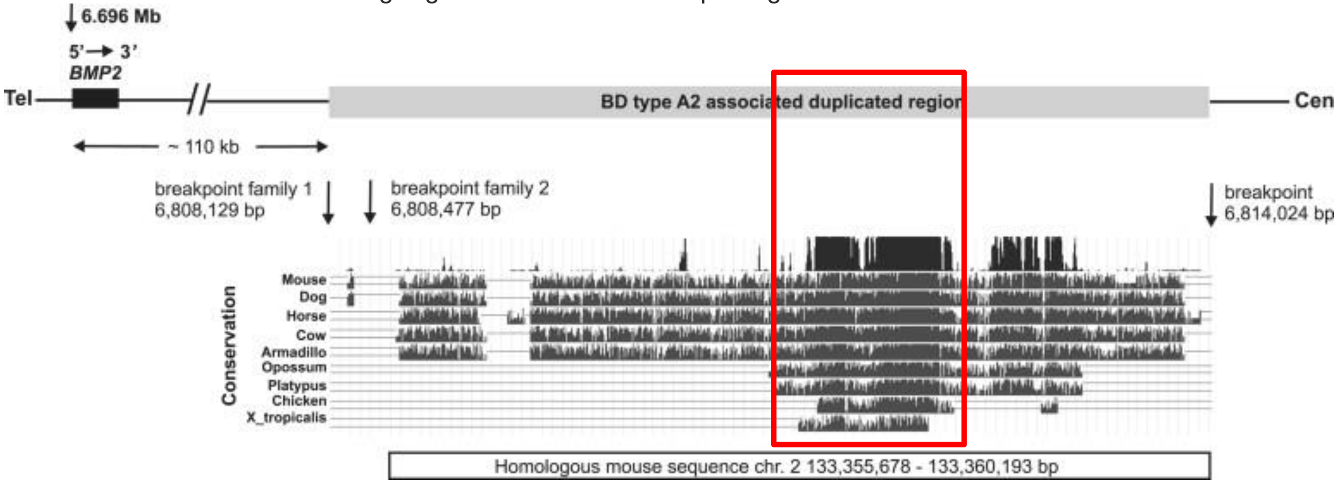
B. Give a major difference between CNV in the human genome versus the mouse genome and comment on how this has changed during evolution of the primate lineage.

CNVs are a lot more common in the human genome, relative to the mouse genome. A lot of nuclear pore genes and neuronal development genes have multiple copies, which is thought to have caused large evolutionary changes in primates.

C. Identify a class of human diseases that is prominently affected by CNV.

Triplicate repeat disorders

Below in figure 1-1 autosomal-dominant brachydactyly type A2 (BDA2) is a limb malformation featured by hypoplastic middle phalanges of the second and fifth fingers. Mutations in genes such as Bone morphogenetic protein receptor 1B (BMPR1B) or Growth and differentiation factor 5 (GDF5) are causes known thus far. These known genes behave according to a simple Mendelian recessive model. However, in another study, duplication of a ~5.5kb region (shown below) was found to be associated with a BDA2 phenotype, demonstrating that Copy Number Variations of non-coding regions can contribute to pathogenesis of human disease.



D. Where is the cis-regulatory module for the above case most likely localized? (use words or draw)

(see red box)

E. In the same study, the authors proposed that the duplication of the cis-regulatory module results in too much *Bmp2* expression, thus deregulating the fine-tuned BMP pathway by competing with another ligand GDF5 and disturbing BMPR1B (one of its receptors) signaling. Propose an alternate mechanism to explain how this mutation could act in a dominant negative manner.

The duplication could result in new interactions between this region and another gene, disrupting the interaction between BMP2. This would result in too little Bmp2 expression, which would adversely affect the BMP pathway as well.

F. A study of global CNVs in human genome has shown that deletions are less likely to associate with RefSeq genes among healthy individuals than duplications. Postulate why.

If an important gene is deleted, than the individual with the deletion mutation will most likely die without passing on the mutation. However, a duplication might not have adverse effects because the additional copy is redundant. Thus, an individual with a duplication is more likely to pass down their mutation.

2. Cancer genomics 12 points

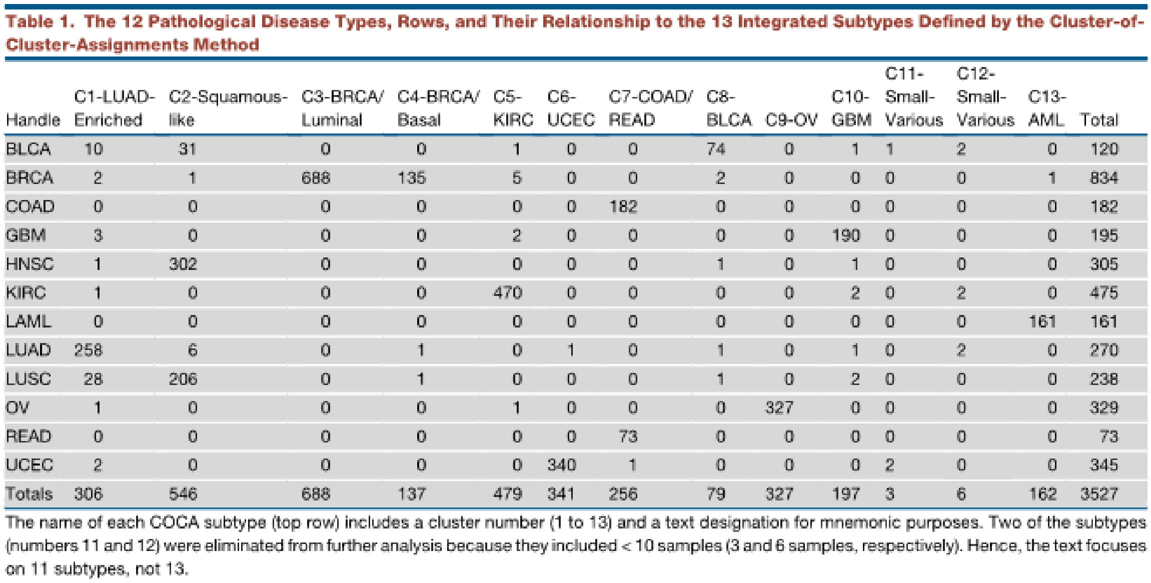


Table 1. The 12 Pathological Disease Types, Rows, and Their Relationship to the 13 Integrated Subtypes Defined by the Cluster-of-Cluster-Assignments Method

Handle	C1-LUAD-Enriched	C2-Squamous-like	C3-BRCA/Luminal	C4-BRCA/Basal	C5-KIRC	C6-UCEC	C7-COAD/READ	C8-BLCA	C9-OV	C10-GBM	C11-Small-Various	C12-Small-Various	C13-AML	Total
BLCA	10	31	0	0	1	0	0	74	0	1	1	2	0	120
BRCA	2	1	688	135	5	0	0	2	0	0	0	0	1	834
COAD	0	0	0	0	0	0	182	0	0	0	0	0	0	182
GBM	3	0	0	0	2	0	0	0	0	190	0	0	0	195
HNSC	1	302	0	0	0	0	0	1	0	1	0	0	0	305
KIRC	1	0	0	0	470	0	0	0	0	2	0	2	0	475
LAML	0	0	0	0	0	0	0	0	0	0	0	0	161	161
LUAD	258	6	0	1	0	1	0	1	0	1	0	2	0	270
LUSC	28	206	0	1	0	0	0	1	0	2	0	0	0	238
OV	1	0	0	0	1	0	0	0	327	0	0	0	0	329
READ	0	0	0	0	0	0	73	0	0	0	0	0	0	73
UCEC	2	0	0	0	0	340	1	0	0	0	2	0	0	345
Totals	306	546	688	137	479	341	256	79	327	197	3	6	162	3527

The name of each COCA subtype (top row) includes a cluster number (1 to 13) and a text designation for mnemonic purposes. Two of the subtypes (numbers 11 and 12) were eliminated from further analysis because they included < 10 samples (3 and 6 samples, respectively). Hence, the text focuses on 11 subtypes, not 13.

Fig 2-1: Classification of cancer samples of different origin histopathologically classified groups (Y-axis) into subgroups (x-axis) according to Cluster of Cluster Assignments (COCA) from multiple genomic measures.

a) What are the top two conclusions YOU draw from the analysis in Fig 2-1, which is taken from the Pan-Cancer study group. Highlight specific relationships/observations to support your point.

Most histopathologically classified groups can be classified under a single COCA subtype. Some cancer samples, like BLCA and BRCA, are classified under multiple COCA subtypes, but most histopathologically classified samples seem to have a dominant COCA subtype. For example, LUSC has several samples under C1-LUAD-Enriched, but an overwhelming majority of the samples fit under the C2-Squamous-like subtype.

However, multiple histopathologically classified cancer samples can be attributed to a single COCA subtype. For example, C2-squamous-like is almost equally probable to be associated with LUSC or HNSC. The C12-small-various subtype is associated with multiple histopath- groups with equal probability.

b) Examine the CNV data in fig 2-2 below, which shows somatic copy-number alterations – red for amplification, blue for deletion – in different chromosomes (Y-axis) across different subgroups of samples (X-axis). Now, comment on the COCA classifications with respect to overall CNV patterns: In doing this, compare in particular C2, C10 and C13 with each other and relative to the entire set.

The samples within a particular COCA subtype have similar copy number variations, which show up as discrete red or blue bands in the data.

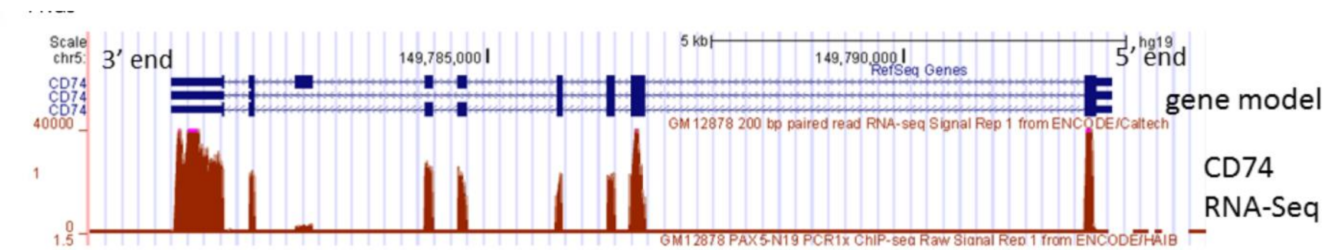
c) What is chromothrypsis and is there a cluster that you think is most likely to have this behavior based on what you see here? Which cluster/tumor type is it and what's your reason?

Chromothrypsis occurs when deletions form extra-chromosomal rings. C10-GBM seems likely to exhibit this behavior because almost every sample displays a large deletion in chromosome 10.

d) You have learned that EGFR is normally associated with signaling cell proliferation in many cell types, so it is not surprising that it can operate as an oncogene. In data of Fig 2-3 shown immediately above, it is altered at a significantly elevated frequency relative to other tumor types in glioblastoma multiforme, (GBM). Based on the data in fig 3-2, what mutational mechanism would you suggest is a major contributor for EGFR in GBM? Explain your answer. If relevant, consult the human genome browser to support your explanation.

Part 3 Gene structure and expression 11 points

Below is a Genome Browser graphic describing the differential regulation of two genes in two different cell types. Cell type 1 is pre-B cell lymphoblastoid (or just B-cells for our purpose), and cell type 2 is HepG2 (or liver for our purpose).


Scale
chr5:
3' end
149,785,000
5 kb
149,790,000
hg19
5' end
RefSeq Genes
CD74
CD74
CD74
gene model
40000
GM 12878 200 bp paired read RNA-seq Signal Rep 1 from ENCODE/Caltech
1
CD74
RNA-Seq
0
1.5
GM12878 PAX5-N19 PCR1x ChIP-seq Raw Signal Rep 1 from ENCODE/HAIB

3a. What do the expression data shown above tell you about splice isoform use for CD74? Be specific about what evidence you are using to make your conclusion. Specify the informative kind of read that you would look for in the primary RNA-Seq data to definitively prove the isoform map you deduced based on read density.

The low read density of the third ChIP-seq peak shows that most CD74 isoforms do not include exon 3.

3b. RNA-seq is usually quantified in RPKM (or the conceptually equivalent FPKM) units. What is the definition of RPKM? If you calculated expression values in RPM instead, how would that distort your comparisons of RNA abundance and what group of genes would be most affected?

RPKM = Reads per kilobase per million

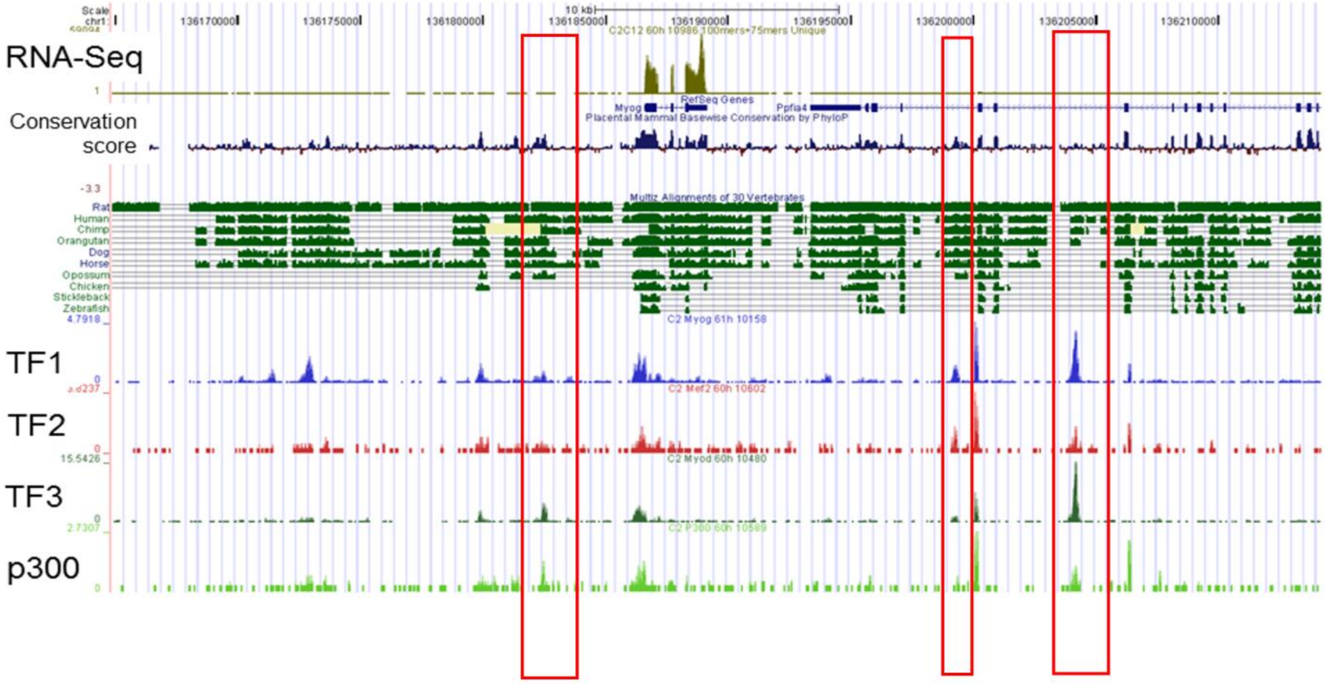
RPM does not take into account the length of the transcript, so smaller genes would appear to be expressed at a lower level than they actually are.

3c. Of the two genes shown, which one is the likely target of the transcriptional regulators shown?

Myog

3d Of the candidate regulatory elements suggested by the data, identify the three you think are most likely to be active transcriptional enhancers in these cells. Briefly, why?

The elements boxed in red are good candidates because they bind to the transcription factors shown and they are enhancers because Myog is expressed (based on the RNA-seq data) when the factors are bound.



3e It is sometimes said that “sequence conservation is king” for highlighting cis-acting regulatory modules (CRM). What do the data at this locus suggest to you about this generality? (you can/should be brief in answering this).

Not all CRMs are conserved. In this example, one of the three elements displays almost no conservation (the rightmost element). Also, there are other conserved regions (to the left of the gene) that do not appear to be CRMs.

> 3F. Now Inspect the list that results and describe which categories are present. Does this make biological sense, considering the tissue samples it is from? Why?

These genes are mainly involved in neuronal and glial development. This makes biological sense, considering that these genes were highly expressed in brain tissue samples.

END