Xiaomi Du
Bi188 Midterm

**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.

Xiaomi Du
Bi188 Midterm

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 in protein coding genes.

**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.**

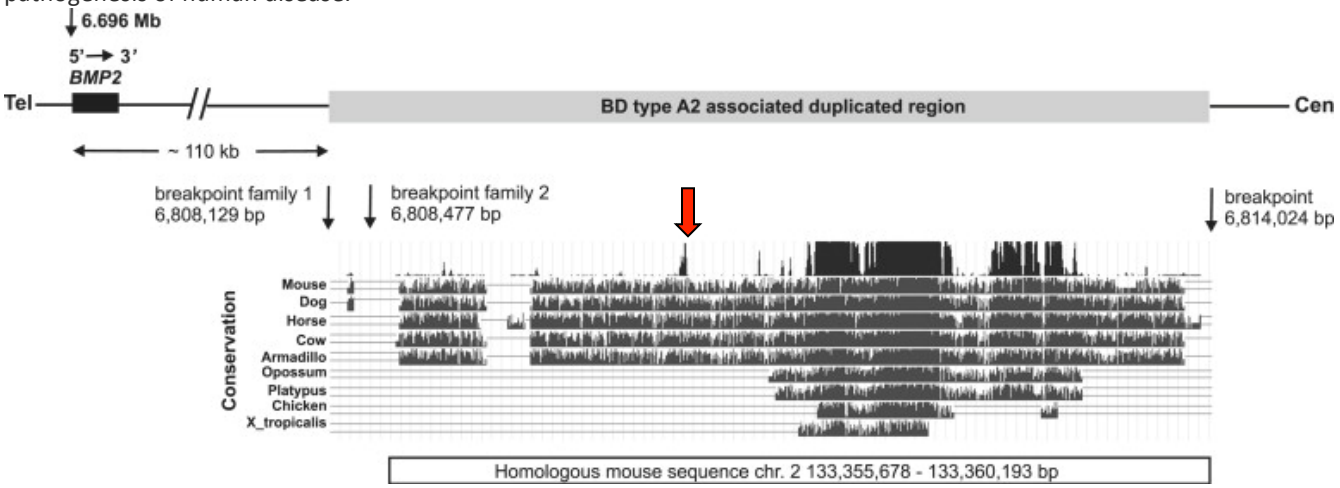
One major difference between CNV in the human genome versus the mouse genome is that the human genome has more duplications, with more diversity in locations, than the mouse genome.

This has changed during evolution of the primate lineage because there was a burst of CNV activity for primates, although orangutans did not have the same burst.

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

One class of human diseases that is prominently affected by CNV includes the autism spectrum.

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)**

The cis-regulatory module for the above case is most likely localized at the peak specified by the red arrow, which is upstream of the rest of the gene.

**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.**

Duplication of the cis-regulatory module could be causing the BDA2 phenotype by overexpressing *Bmp2* and disrupting GDF5-BMPR1B signaling. However, a dominant negative mutation is defined as one that produces a worse phenotype with a wild-type allele than the wild-type allele paired with a null allele. Thus, this mutation could be acting in a dominant negative manner in a multitude of ways.

One alternate mechanism for how this duplication can act in a dominant negative manner could be that the excess Bmp2 binds to the BMPR1B receptor without falling off, thus constitutively preventing GDF5 from binding and activating the receptor. Alternatively, the mutation could be duplicating a binding site for a microRNA against the GDF5 ligand or BMPR1B.

**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.**

Deletions might be less likely to associate with RefSeq genes among healthy individuals than duplications because RefSeq may try to associate the sequence formed by the deletion with a normal gene.

After there is a deletion in a gene, the sequences up and downstream of the deletion are joined together. If RefSeq does not recognize the two separate ends of the deletion as separate parts, it might try to align the newly formed sequence with a wild-type sequence. However, this attempt will not be fruitful, and it could hinder the rest of the alignment.

On the other hand, a duplication event is more likely to associate with RefSeq genes because it has the wild-type gene in it; RefSeq needs to disregard the duplication, which is easier than deciphering how to separate a sequence. Therefore, it is more likely for duplications to associate with RefSeq genes than for deletions.

Xiaomi Du
Bi188 Midterm

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.**

The top two conclusions I draw from the analysis in Figure 2-1 are:

1) Each histopathologically originating cancer sample correlates strongly with one subgroup, as identified by COCA from multiple genomic measures. Thus, each cancer type is caused by mutations in predominantly one subgroup.

2) However, some subgroups correspond to more than one type of cancer sample. Thus, the subgroups can cause more than one kind of cancer.

The first conclusion can be seen by comparing the total number of each kind of cancer sample with the individual values of that cancer sample for each subgroup. For example, there are 161 samples under AML for LAML, and there are 161 total LAML samples. Therefore, for the LAML samples observed in the Pan-Cancer study group, all were caused by mutations in the AML subgroup. Similar observations can be seen for COAD and COAD/READ, for all the Y-axis groups, One subgroup: C3-6, C8-13

More than one: C1-2, C7

**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 COCA classifications are consistent in their overall CNV patterns. For example, almost all of the cancer samples in C13 barely have any CNV activity; their somatic copy numbers are mostly wild type. On the other hand, the majority of the tumor samples in C10 all have strongly defined amplification of chromosome 7 and deletion of chromosome 10, with weaker expression of amplifications and deletions in the other chromosomes. Lastly, even though the tumors in C2 exhibit many mutations, and there is a random pattern of amplifications

and deletions throughout the chromosomes, all the tumors have very similar genotypes; they all have deletions in the top half of 1, amplifications in the bottom half of 1 and all of 2, deletions in half of 3, and so on.

The rest of the subgroups are similarly uniform within each subgroup. C9, C4, C1, C8, C3, and C5 resemble C2 in their abundant switches from amplifications to deletions, and all of their tumors still have consistent expression within each subgroup. Likewise, C7 takes after C10 in having a few prominent CNVs, while C6 has a bit more CNVs than C13, but they are very mild. For all of the subgroups, the tumors classified within the subgroup have the same CNV patterns.

Many subgroups have similar amplifications or deletions. For example, C2, C7, C8, and C10 all have duplications of chromosome 7. However, they are each different in their expressions of other chromosomes.

**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 is the complete breakage and rearrangement of a single chromosome. A cluster that is most likely to have this behavior would be C10?

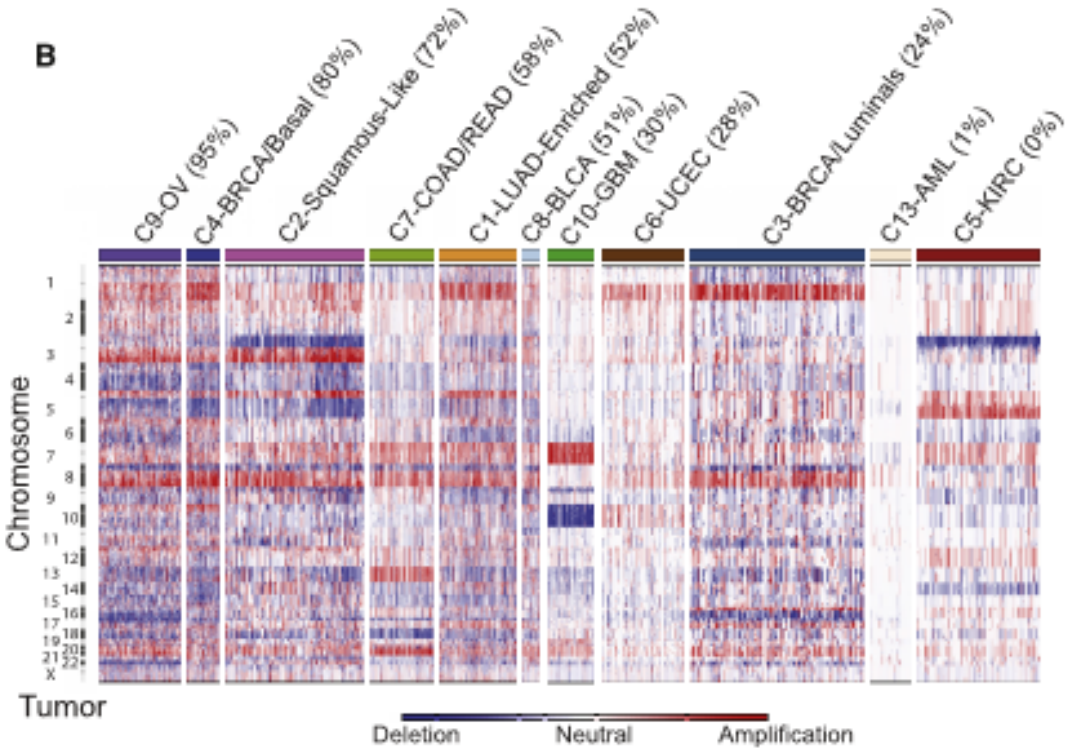



Figure 2-2. Somatic copy-number alterations in clustered tumors (X-axis) on different chromosomes (Y-axis). Red shows amplification, blue shows deletion.

Xiaomi Du
Bi188 Midterm

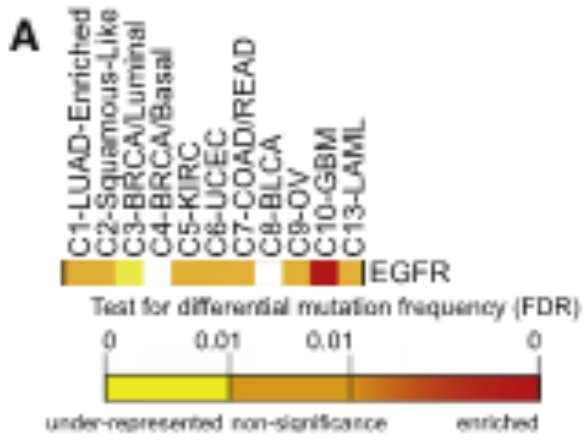



Figure 2-3. Mutation rate of EGFR, a gene associated with cell proliferation, across different COCA subtypes.

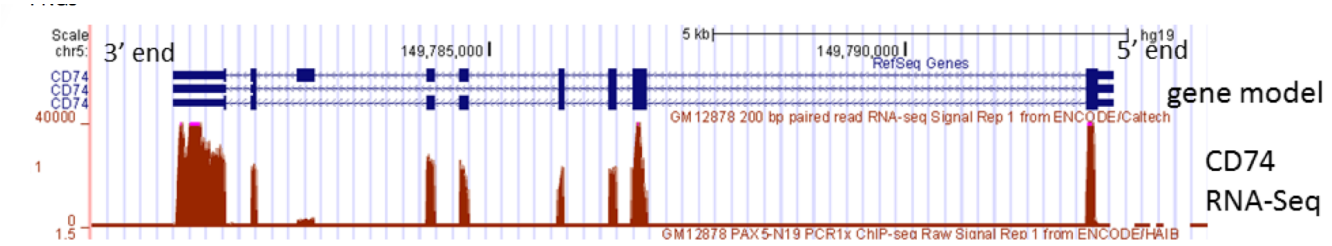
d) You have learned that EGFR is normally associated with signaling cell proliferation in many cell types, so it is not surprising that it can operate as an oncogene. In data of Fig 2-3 shown immediately above, it is altered at a significantly elevated frequency relative to other tumor types in glioblastoma multiforme, (GBM). Based on the data in fig 2-3, 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.

A mutational mechanism that is a major contributor for EGFR in GBM is chromothrypsis. The frequency of mutation is so enriched in GBM, especially in comparison to the other subgroups. According to the human genome browser, EGFR is located at the end of chromosome 7. The GBM subgroup has significant amplification of chromosome 7, according to Figure 2-2.

Xiaomi Du
Bi188 Midterm

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).




**3a. What does 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 splice isoform use for CD74 is pretty consistent because the RNA-Seq peaks match the exons of the genes exactly. Therefore, there is one isoform of CD74. You would look for RPKM in primary data.

**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 stands for *r*eads *p*er *k*ilobase-*m*inute. If we had calculated expression values in RPM instead, that would distort our comparisons of RNA abundance because it would not normalize for the number of base pairs read; this would affect the promoter genes the most.

3c. In the figure below a locus containing two genes is shown together with RNA-Seq data and ChIP-seq data. All measurements were done from the same cell preparation, for three pertinent transcription factors and the p300 histone acetyltransferase enhancer protein. Information on sequence conservation appears in the summary track and is shown in further detail in the alignment plots from species as distant as fish. Annotate the figure as you wish to aid in answering the questions (you can drop the png into ppt and do it that way; use any other program you like; or go old school and print it to draw on). It is OK to draw first and then transfer after exam in over to some electronic form to send – but the answer itself cannot be changed when you do that).

**Of the two genes shown, which one is the likely target of the transcriptional regulators shown?**

The gene denoted by the black rectangle is the likely target of the regulators. It expresses RNA, and its 5' end shows great homology with the other species' genes.

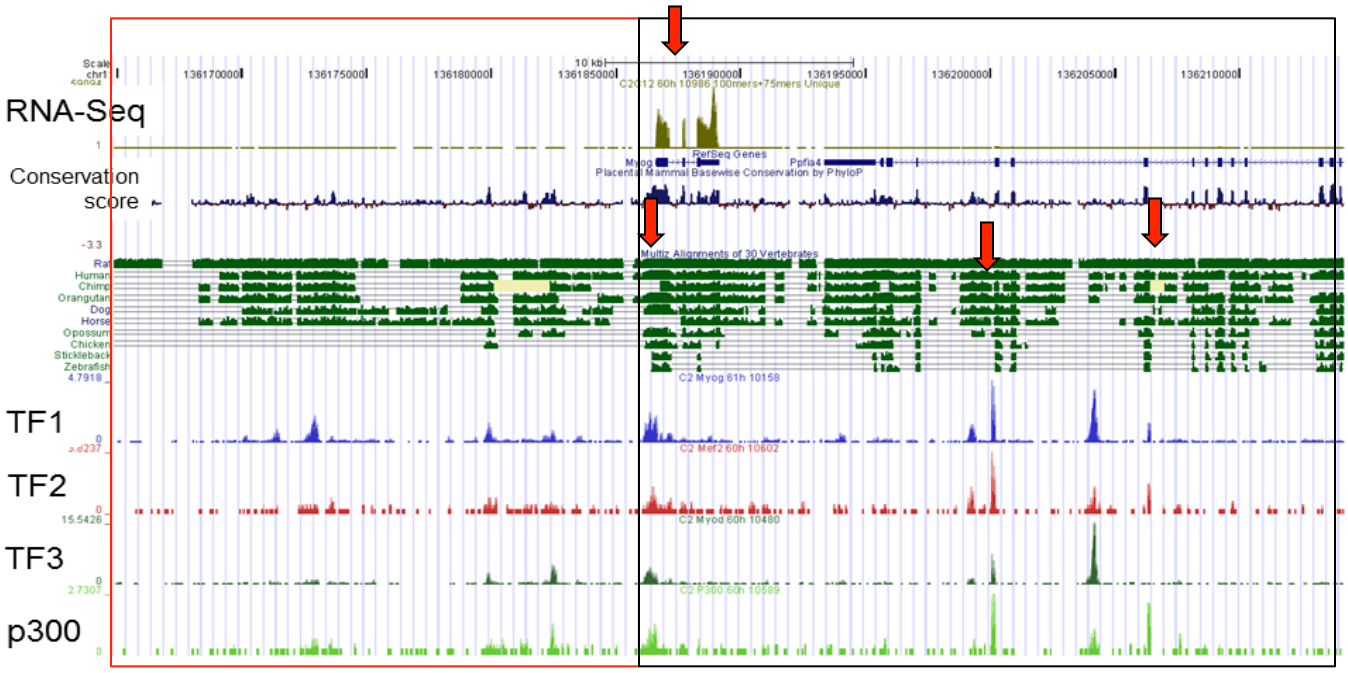
**3d. Of the candidate regulatory elements suggested by the data, identify the three you think are most likely to be active transcriptional enhancers in these cells. Briefly, why?**

Xiaomi Du
Bi188 Midterm

TF1, TF2, and p300 are likely to be active transcriptional enhancers in these cells because they act on the most conserved sequences.

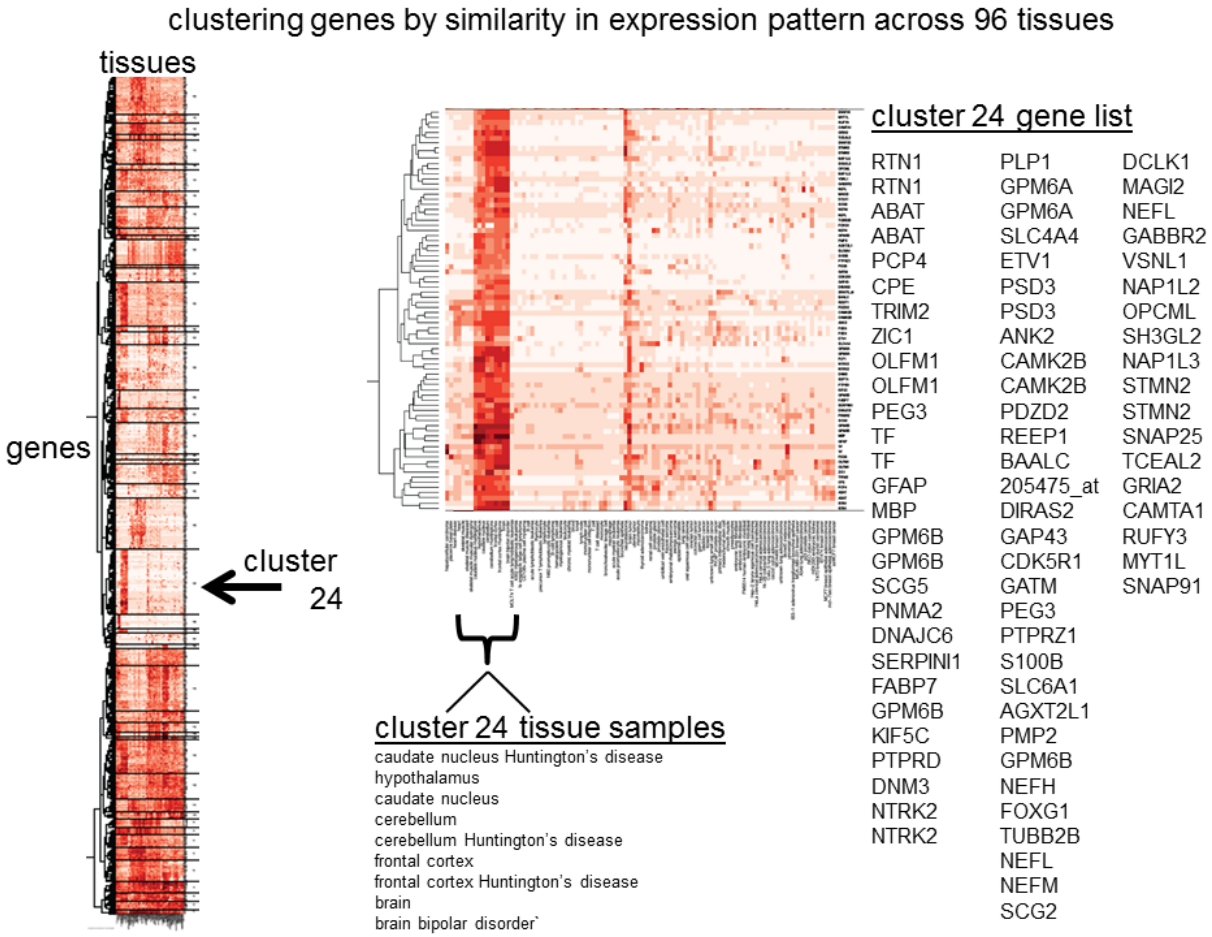
**3e. It is sometimes said that "sequence conservation is king" for highlighting cis-acting regulatory modules (CRM). What do the data at this locus suggest to you about this generality? (you can/should be brief in answering this).**

The data at this locus supports this generality because the transcription factors and p300 act mostly on heavily conserved areas, except for the one series of peaks under 136205000, in which only 4 of the animals show conservation.



Xiaomi Du
Bi188 Midterm

Below, you see transcriptomes clustered from a collection of 96 tissues that were assayed for RNA using a microarray platform. 14,000 genes were quantified. A 2-way hierarchical clustering was used to group the genes by similarity of expression profile across the tissues, and to group the tissues by similarity of genes expressed in them.




Focus on cluster 24 (arrow on left side). You can see that this cluster of 9 samples (see middle callout box) is very well defined from the other tissues in the study. The gene expression pattern (red is high, white is low) is very distinct in this cluster. The genes in this cluster are called out on the right side of the panel.

Follow these instructions to answer the question at step (6):

1)Go to this website: (you are allowed to use it for the rest of this question)
http://llama.mshri.on.ca/funcassociate/

2) in the dialog box for species (upper left), select Homo sapiens

3) in the “choose a namespace” dialog box, select “hgnc_symbol”

4) “Provide a list of genes as a query”, copy and paste the list of gene names below:

RTN1
RTN1
ABAT
ABAT
PCP4
CPE
TRIM2

Xiaomi Du
Bi188 Midterm

ZIC1
OLFM1
OLFM1
PEG3
TF
TF
GFAP
MBP
GPM6B
GPM6B
SCG5
PNMA2
DNAJC6
SERPINI1
FABP7
GPM6B
KIF5C
PTPRD
DNM3
NTRK2
NTRK2
PLP1
GPM6A
GPM6A
SLC4A4
ETV1
PSD3
PSD3
ANK2
CAMK2B
CAMK2B
PDZD2
REEP1
BAALC
205475_at
DIRAS2
GAP43
CDK5R1
GATM
PEG3
PTPRZ1
S100B
SLC6A1

AGXT2L1
PMP2
GPM6B
NEFH
FOXG1
TUBB2B
NEFL
NEFM
SCG2
DCLK1
MAGI2
NEFL
GABBR2
VSNL1
NAP1L2
OPCML
SH3GL2
NAP1L3
STMN2
STMN2
SNAP25
TCEAL2
GRIA2
CAMTA1
RUFY3
MYT1L
SNAP91

5) Click the "Functionate" button.

**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?**

The categories present include genes with attributes related to the nervous system, such as "neurofilament bundle assembly," "neuron recognition," "central nervous system development," "axon," "synapse," and "cell projection." This makes sense at first, given the tissues samples it is from, because they are all neural. However, Huntington's is a neurodegenerative disease, and these tissue samples compare normal tissues to Huntington or bipolar disorder tissues; there is cerebellum tissue and Huntington's disease cerebellum tissue, for example. However, upon considering that the Huntington tissue should not have the same expression of neural genes as normal tissue, this pattern does not make biological sense.

END