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.

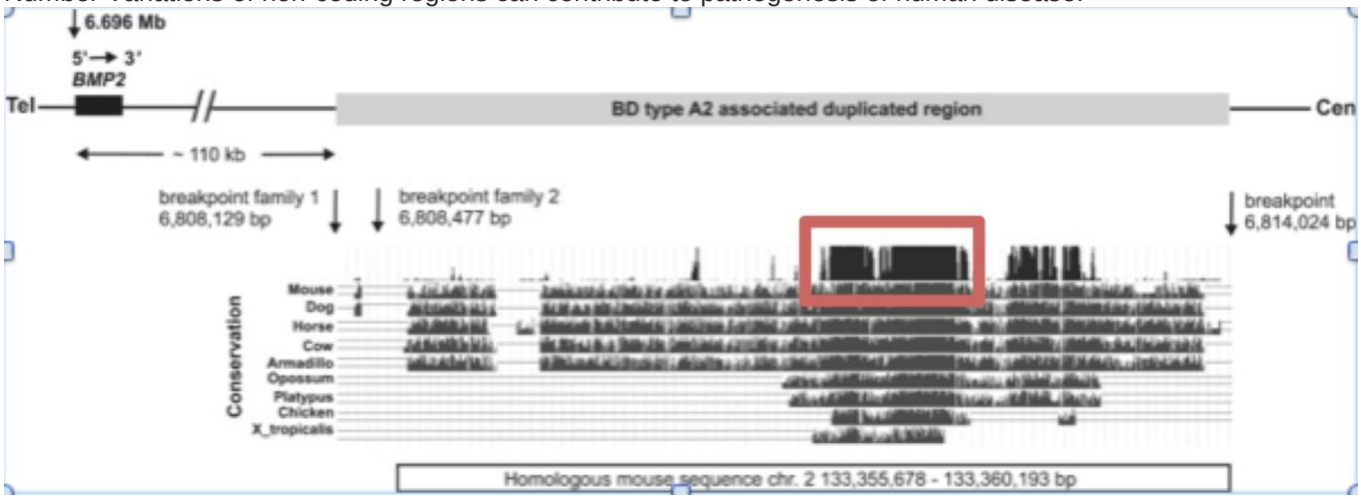
1. Genome structure and Copy Number Variation (total of 12 points)

A. Where are Copy Number Variations (CNVs) more likely to occur in the human genome?
CNVs are more likely to occur in centromere and telomere regions of the chromosomes.

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.
Mouse CNVs are tandem and local, and they do not usually move between chromosomes. In contrast, human CNVs are more interspersed and global, and they move around multiple chromosomes. During evolution of the primate lineage, there was a burst of CNVs when the primate lineage separates into orangutans and common ancestors of gorillas, chimpanzees, and humans (the burst was observed in transition to the latter).

C. Identify a class of human diseases that is prominently affected by CNV.
Most CNVs result in mental disorders that related to brain functioning such as autism and schizophrenia.

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 is likely to be contained in the most conserved region, as shown in red box above.

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.

BMP2 may bind to another type of BMP receptor, and the resulting secondary signaling competes with BMPR1B signaling. Overexpression of BMP2 causes an imbalance between these two signaling pathways by increasing the secondary signaling and decreasing the normal BMPR1B signaling.

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 lead to the loss of gene function while duplication leads to the overexpression of gene function. In general, loss of gene function is phenotypically more severe than overexpression of gene function, and the loss of gene function from deletions is thus more likely to cause related diseases.

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.

(1) There are cases when different cancer groups share the same subgroups of COCA classifications. For example, BLCA, LUAD, and LUSC cancer groups share the common C1 subgroup; also, BCLA, HNSC, LUSC, and relatively few LUAD share the common C2.

(2) Certain cancer groups have multiple subgroups. For example, BCLA cancer group has C1, C2, and C3 subgroups; also, BRCA has C3 and C4; lastly, LUSC has C1 and C2.

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.

Each subgroup of COCA classifications has a unique CNV profile. Subgroups C9, C4, C2, C1, C8, and C3 (note the inclusion of C2) suffer severe somatic CNVs prevailing throughout the whole genome of chromosomes. Subgroups C7 and C5 suffer moderate somatic CNVs throughout the whole genome. C6 suffer even milder somatic CNVs, and C13 lacks any significant somatic CNVs. C10 has the somatic CNVs particularly at chromosome 7 (amplification of the whole chromosome) and chromosome 10 (deletion of the whole chromosome). C10 also has a portion of its chromosome 9 deleted.

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 process of chromosome(s) getting broken into pieces and recombined in random order with some broken piece(s) lost in the process. Consider C10-GBM, which has a portion of its chromosome 9 deleted (represented by a blue band). This partial deletion may be result of some lost pieces from chromothrypsis in chromosome 9, and C10 is thus the most likely cluster that exhibits the behavior of chromothrypsis.

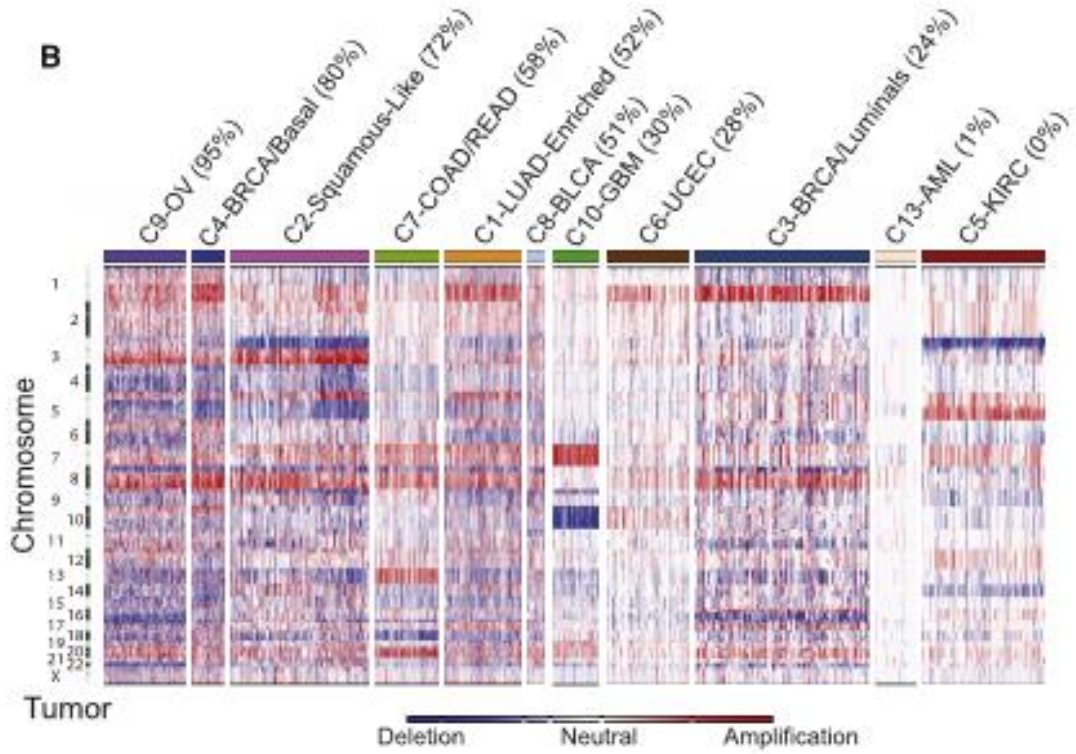


Figure 2-2. Somatic copy-number alterations in clustered tumors (X-axis) on different chromosomes (Y-axis). Red shows amplification, blue shows deletion.

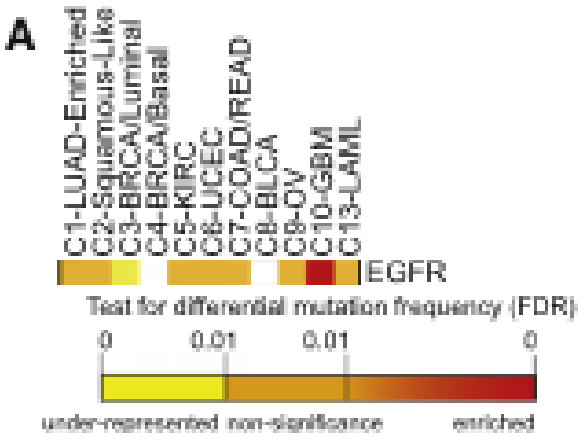


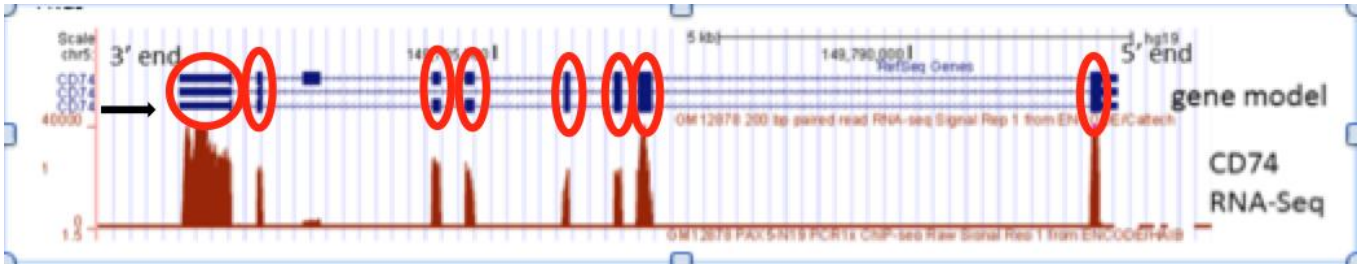
Figure 2-3. Mutation rate of EGFR, a gene associated with cell proliferation, across different COCA subtypes.

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 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.

Human genome browser indicates that EGFR gene is located in chromosome 7. Figure 2-2 shows that chromosome 7 in GBM has undergone copy-number amplification, and overexpression of EGFR (whose copy-number has also undergone amplification) is expected as a result. Overexpression of EGFR induces an increase in signaling cell proliferation, which eventually results in cancer-type phenotypes of GBM.

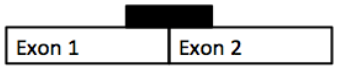
Part 3 Gene structure and expression 11 points

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 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.

RNA-Seq data show which exons of a given gene are expressed in the mRNA transcript. If a certain exon is expressed, we should expect a density peak of reads in the RNA-Seq data. According to the given CD74 RNA-Seq data in the above figure, the exons in the circled regions need to be expressed. The isoform that corresponds to these RNA-seq data is the one in the third row (indictated by the arrow on the left). To prove this, I need reads that contain both portions of neighboring exons. The shaded rectangle shown in the figure below is such type of reads that is needed. Note that the two exons drawn below are after the splice junction between them is removed by alternate splicing.

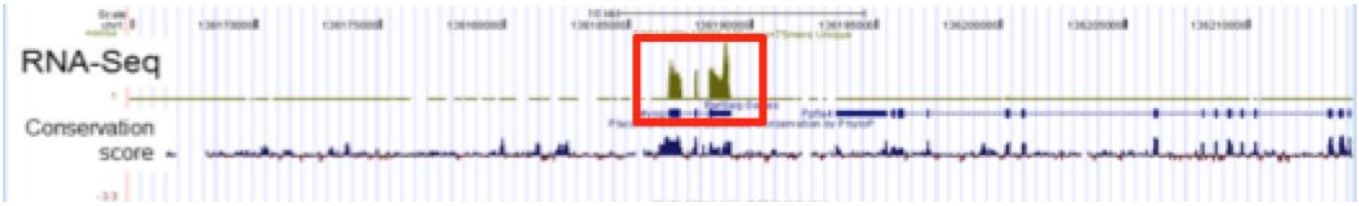


3b. RNA-seq is usually quantified in RPKM (or the conceptually equivalent FPKM) units. What is the definition of RPKM? If you calculated expression values in RPM instead, how would that distort your comparisons of RNA abundance and what group of genes would be most affected?

By definition, RPKM is the Reads Per Kilobase of exon per Million reads, and it is an unit to normalize the reads according to the depth of sequencing and length of the transcript of a gene. If we were to use the RPM instead, without normalizing the reads according to the transcript's length, genes with longer transcript would be most affected. By probability rules, longer transcripts are likely to get more reads than shorter transcripts, and RPM thus gives the misleading data in which genes with longer transcripts come out in greater abundance than they actually are in reality.

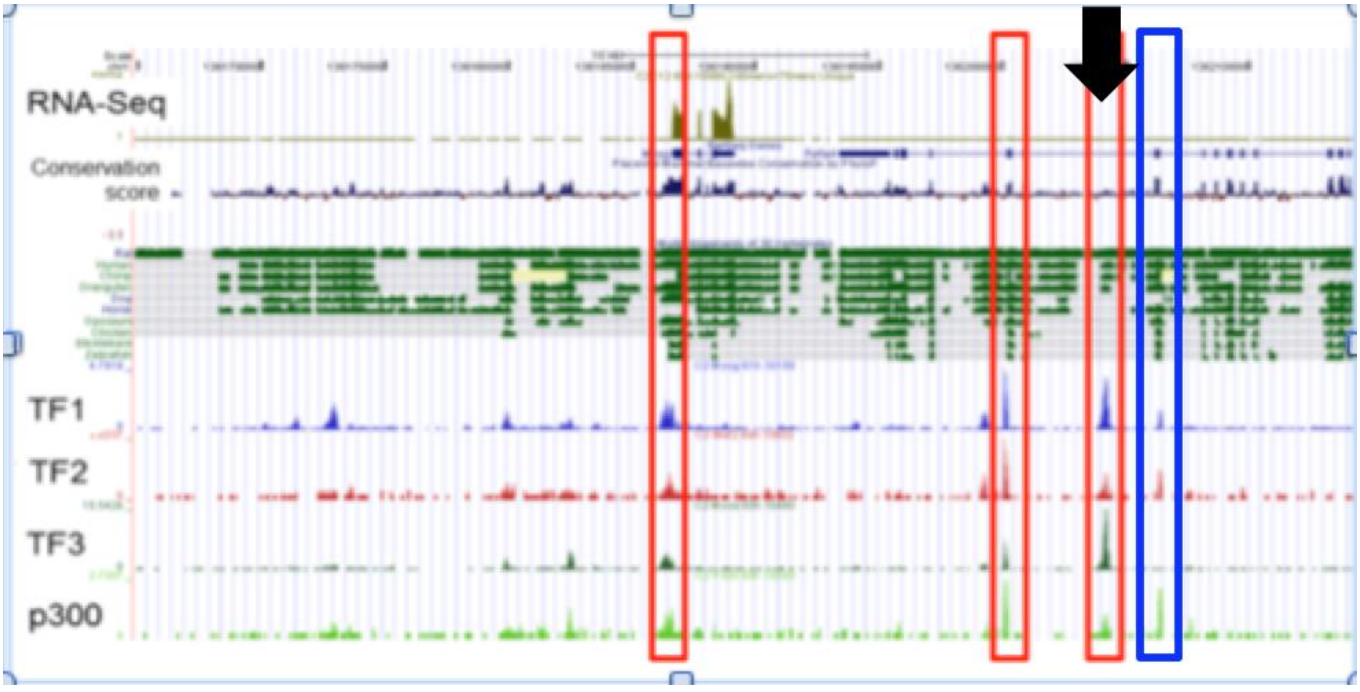
3c In the figure below a locus containing two genes is shown together with RNA-Seq data and ChIP-seq data. All measurements were done from the same cell preparation, for three pertinent transcription factors and the p300 histone acetyltransferase enhancer protein. Information on sequence conservation 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?



As shown on the RNA-Seq data, only Myog is expressed into mRNA transcript, and Myog is thus the likely target of the transcriptional regulators.

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?

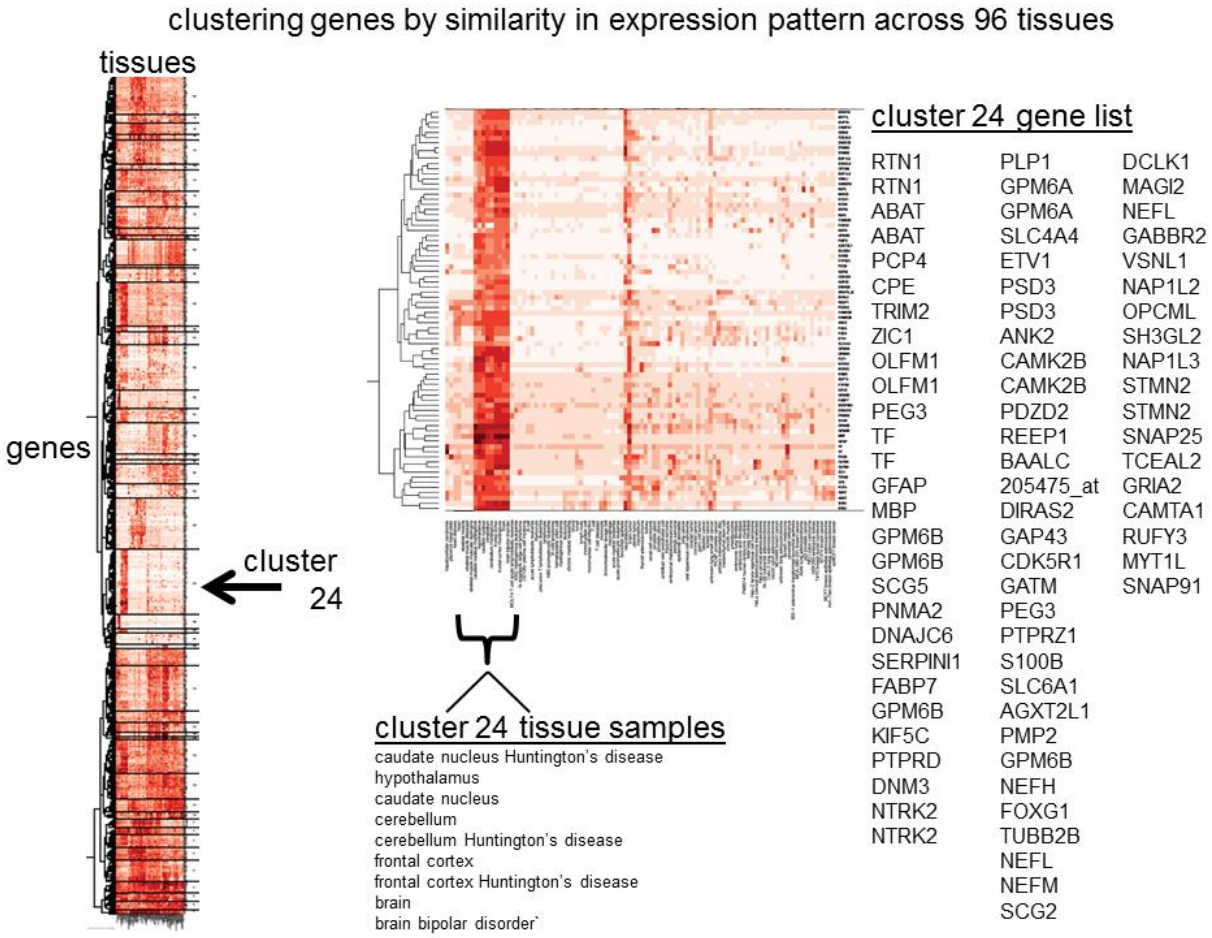


The most likely transcriptional enhancer regions would have density peaks of TF1, TF2, TF3, and p300 from ChIP-Seq data (i.e. these four proteins bind to the enhancer region). Three regions are boxed in red in the figure above. Note the possibility that the first red-boxed region may be a promoter region. If we assume the region to be a promoter, the next candidate for an enhancer region is boxed in blue.

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).

Note that the third red-boxed region (indicated by a shaded arrow) in the figure from 3d, which is one of the most likely cis-regulatory modules predicted by ChIP-Seq, is not conserved. Thus, the data at this locus suggest that the generality (that “sequence conservation is king” for highlighting CRM) does not always hold true.

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:

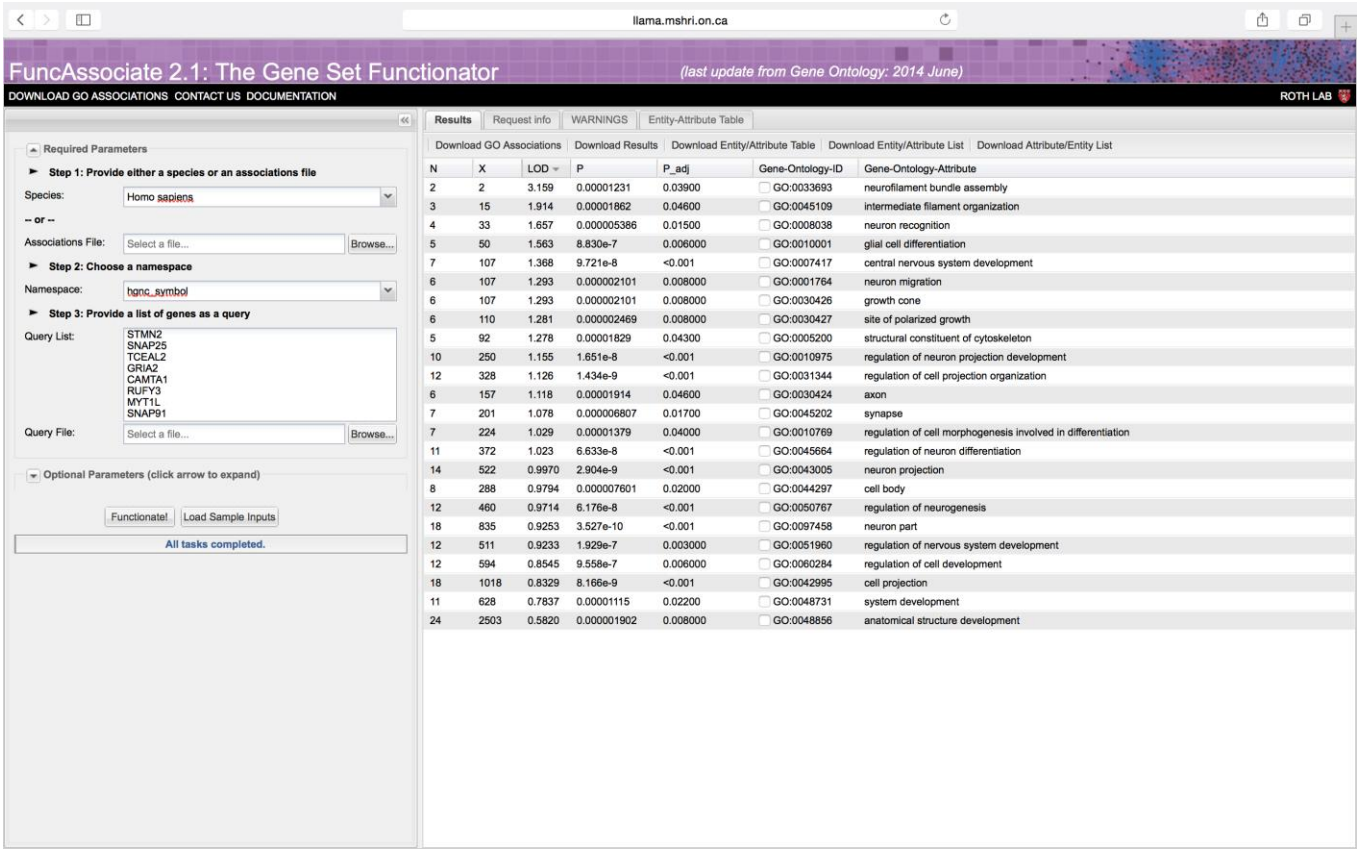
RTN1
RTN1
ABAT
ABAT
PCP4
CPE
TRIM2
ZIC1
OLFM1
OLFM1
PEG3
TF
TF
GFAP
MBP
GPM6B
GPM6B
SCG5
PNMA2
DNAJC6
SERPINI1
FABP7
GPM6B
KIF5C
PTPRD
DNM3
NTRK2
NTRK2
PLP1
GPM6A

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?



N	X	LOD	P	P_adj	Gene-Ontology-ID	Gene-Ontology-Attribute
2	2	3.159	0.00001231	0.03900	GO:0033693	neurofilament bundle assembly
3	15	1.914	0.00001862	0.04600	GO:0045109	intermediate filament organization
4	33	1.657	0.000005386	0.01500	GO:0008038	neuron recognition
5	50	1.563	8.830e-7	0.006000	GO:0010001	glial cell differentiation
7	107	1.368	9.721e-8	<0.001	GO:0007417	central nervous system development
6	107	1.293	0.000002101	0.008000	GO:0001764	neuron migration
6	107	1.293	0.000002101	0.008000	GO:0030426	growth cone
6	110	1.281	0.000002469	0.008000	GO:0030427	site of polarized growth
5	92	1.278	0.00001829	0.04300	GO:0005200	structural constituent of cytoskeleton
10	250	1.155	1.651e-8	<0.001	GO:0010975	regulation of neuron projection development
12	328	1.126	1.434e-9	<0.001	GO:0031344	regulation of cell projection organization
6	157	1.118	0.00001914	0.04600	GO:0030424	axon
7	201	1.078	0.000006807	0.01700	GO:0045202	synapse
7	224	1.029	0.00001379	0.04000	GO:0010769	regulation of cell morphogenesis involved in differentiation
11	372	1.023	6.633e-8	<0.001	GO:0045664	regulation of neuron differentiation
14	522	0.9970	2.904e-9	<0.001	GO:0043005	neuron projection
8	288	0.9794	0.000007601	0.02000	GO:0044297	cell body
12	460	0.9714	6.176e-8	<0.001	GO:0050767	regulation of neurogenesis
18	835	0.9253	3.527e-10	<0.001	GO:0097458	neuron part
12	511	0.9233	1.929e-7	0.003000	GO:0051960	regulation of nervous system development
12	594	0.8545	9.558e-7	0.006000	GO:0060284	regulation of cell development
18	1018	0.8329	8.166e-9	<0.001	GO:0042995	cell projection
11	628	0.7837	0.00001115	0.02200	GO:0048731	system development
24	2503	0.5820	0.000001902	0.008000	GO:0048856	anatomical structure development

Note that most of the gene-ontology-attributes (e.g. axon, neural part) are related to neural tissue. In other words, genes involved in constructing neural tissues are highly expressed in the 9 tissue samples. This makes biological sense, because the 9 tissue samples used in this case are indeed neural tissues in origin.

END