**BIOM262 Final Exam**

**Part 1 – DNA Analysis and Variant Calling (Harismendy):**

Please include the code you used to generate the result in your answer. The datasets are the same ones used in class. HINT – start an interactive job and run all your commands there. Don’t submit scripts.

1. Which CGC gene has the shortest exon? What is its length?
2. Which CGC gene has the most exons? How many exons?
3. What fraction of reads from SRR948508 sample align to chromsome 1? Provide numerator and denominator.
4. In the CPTRES4.chr21.bam file, what fraction of the CGC genes base pairs are covered by more than 20 reads?
5. In the CPTRES4.chr21.bam, how many RUNX1 base pairs are covered at 20x or greater?
6. In the CPTRES1vs15.vcf.gz how many non-synonymous variants that PASS filter are common (INFO/SS=1) between control and tumors?

**Part 2 – (Schork):**

\*\*\*WILL BE ADDED LATER\*\*\*

**Part 3 – Answer ONE of the following three questions:**

1. Alon Goren (CHIP-Seq):

* Data files needed to answer this question are located here: /projects/ps-yeolab/biom262\_2017/final\_exam/goren\_chip/

HepG2 is a Human cell line derived from the liver tissue of a male with well-differentiated hepatocellular carcinoma, and is used in many studies providing a model for hepatocytes. The directory has bam files of a single chromosome’s ChIP-seq data for several transcription factors (TFs) and the WCE control, aligned to hg19.

Please provide the code you are using to answer the questions below.

**1.**     What chromosome are the datasets originated from? How did you learn that?

**2.**     What are the differences in the binding of the 3 TFs? Include the following:

a.     Number of binding sites

b.     Estimated IP efficiency

c.     Percent of GC

d.     Estimated fragment length

e.     Enrichment for specific motifs (top 5)

**3.**     Separate the peaks of each dataset between the ones associated with promoters and ones that are not.

a.     How many did you get for each TF in each category?

b.     What datasets would you need to further corroborate the your annotations of the peaks between promoters and enhancers?

**4.**     Finally, make a list of the 5 genes that have the strongest binding by each TF. Do they overlap? Does these genes make sense given the type of cells HepG2 are?

1. Graham McVicker (GWAS/eQTLs):

See the file uploaded to the website with the plots you need to answer this question (McVicker\_Final\_Exam\_Question).

Q1: Under the null hypothesis, how should the p-values be distributed (i.e. what distribution should they come from)?

Q2: Do you notice a problem with the p-values from this plot?

Q3: Why is this a problem?

Q4: What could cause this problem with the p-values?

Q5: How could you correct this problem?

1. Hannah Carter (Networks):

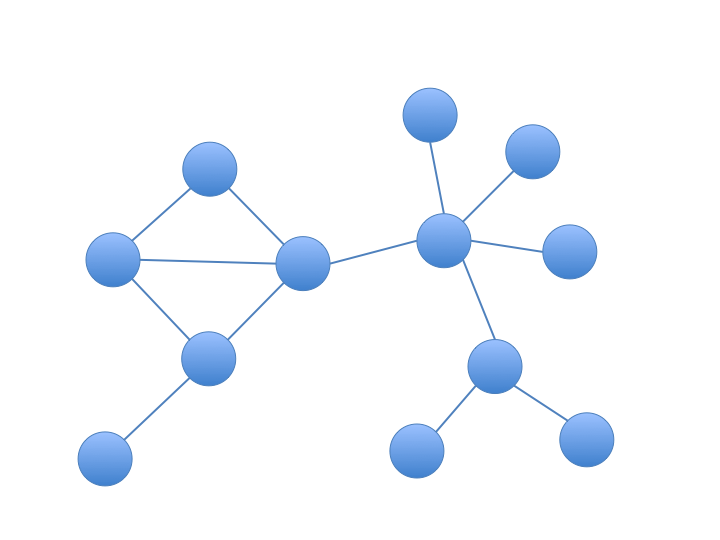
1

1) What is the advantage of a scale-free network architecture? Why might this be a common architecture for biological systems to adopt?

2) What does the clustering coefficient of a node tell you? What are the implications of a node with a high clustering coefficient in a protein-protein interaction network? What about a node with a high clustering coefficient in a gene regulatory network?

3) What kind of information is displayed in a genetic interaction network? What is one application of studying genetic interaction networks?

4) What is the diameter of the following graph? What is the highest degree of a node in the graph?



5) What is ‘degree-preserving’ permutation, and why would you use it instead of simply shuffling the edges of a network?

**SUBMIT!**

The final is due on Sunday March 26 at 11:59 PM. Submit this file in an email to Emily and Jamison with the filename: lastname\_firstname\_final\_exam

Office Hours:

Emily – Monday March 20, 9-10AM (SCRM 3318)

Jamison – Friday March 24, 1:30 – 3:30 (JCVI)