**BIOM262 Final Exam**

IMPORTANT: Please read the instructions first!

This final covers the class material treated since the midterm. Each section here covers a module from class. The first three sections are mandatory (part 1 -3) along with either part 4 or part 5.

Please place the answer of each section into SEPARATE documents. In the end you should have 4 different files which you will be sending to the email specified for each problem.

Each email should have the format: CMM262 Final ‘Section last name’ ‘Student Name’.

Each file should be of the format: cmm262\_student\_name\_sectionLastName

You will lose points if you don’t do so!

The final is due on **Saturday, March 24 at midnight**.

This is a take home final and collaboration or exchanging of ideas is NOT allowed!

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

Please include the code you used to generate the result in your answer. The datasets are found at: /oasis/tscc/scratch/biom200/cmm262/Final\_data\_Harismendy . HINT – start an interactive job and run all your commands there. Don’t submit scripts.

Submit answers to: pfiaux@ucsd.edu

1. Which CGC gene has the shortest exon? What is its length? Files needed: CGC.exons.bed
2. Which CGC gene has the most exons? How many exons? Files needed: CGC.exons.bed
3. What fraction of reads from SRR948508 sample align to chromsome 1? Provide numerator and denominator. Files needed: chr1.fa.gz, SRR948508\_1.fastq.gz, SRR948508\_2.fastq.gz To answer this first obtain the bwa alignement software (conda install -c bioconda bwa) and align using ‘bwa mem’. (Hint 1: use ‘bwa mem’ as a first step, default parameters are sufficient, indexing is not necessary since we already did this for you. Hint 2: samtools has a good function for calculating summary statistics)
4. In the CPTRES4.chr21.bam file, what fraction of the CGC genes base pairs are covered by more than 50 reads? (Hint: look at the samtools depth command and the available options)

**Part 2 – Population Genetics (Nick Schork):**

Use leptospirosis data in /oasis/tscc/scratch/biom200/cmm262/Final\_data\_Schork\_Libiger/lepto\_unrel\*

Include the commands you used to get the answer in your responses.

Submit answers to: pfiaux@ucsd.edu

1) prune for LD in plink and create IBS distance matrix

2) Run multidimensional scaling (MDS) in R on the resulting distance matrix. Plot the individuals using the first 2 components in two-dimensional plot in green along with the in-class generated data for chromosome 21 of the hapma3 project (if you followed along in class you should already have this data set and simply read it into R). Show the plot. (Hint: the MDS result for the leptospirosis data and the hapmap3 data can simply be combined in R. Think of what data you would add to what in order to leverage the row information we used in class to display the different groups)

3) Are there apparent clusters in the plot? What do you think they represent?

4) Do you see evidence of admixture? How is it manifested in the plot?

5) If you were to carry out a genome-wide association study using this data, would you need to consider population stratification? If yes, why? What, if any, strategies would you incorporate in your analysis to account for population stratification? How would you to that using plink (just show what flags and why you would use them, you don’t have to actually run anything here)

**Part 3 - GWAS/eQTLs (Graham McVicker):**

See the file uploaded to the Github website in the ‘Final’ folder with the plots you need to answer this question (BIOM262\_final\_exam\_mcvicker).

Submit answers to: pfiaux@ucsd.edu

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?

**Answer ONE of the following two Sections:**

**Part 4 – ChIP-seq (Alon Goren):**

Submit answers to: pfiaux@ucsd.edu

1. You are interested in understanding the role of a specific genetic dysregulation in cancer. You are focused on two types of samples: case and control. You assume that some genetic dysregulation is causal, thus you want to evaluate and compare putative regulatory regions.

How would you experimentally identify the putative regulatory elements? (provide details describing the type of method you would use, the specific molecular marks you will study, and the reason you chose these marks)

1. Following up on the data you generated above, you analyze the putative regulatory regions and identify a putative enhancer that is present only in the case group and devoid in the controls. Upon further investigation you see that a MYB binding site was created in the case group.   
     
   Describe how you would computationally obtain the regulatory regions from the data you generated in question 2. What is the function of MYB? Based on these properties of MYB, what phenotype do you think the case group has compared to the control group?
2. Your colleagues sent you a BAM file of ChIP-seq data for the transcription factor MYB (/home/ucsd-train36/biom262-2018/exam-data/human\_tall-myb\_chipseq.bam) in human T-All cells, but you suspect that during the process of generating the file, they did something wrong.
   1. Can you identify the issue? (hint – use samtools and view the file using the flag to display the header too).
   2. Now that you know what went wrong, how would you fix the problematic BAM file using the ‘bedtools’ command ‘bamtofastq’, and other tools that were introduced in class (no need to do so, just spell out the steps you would do)?
3. Next, you want to see with what other genomic regions the new MYB binding site interacts. What experiment would you perform to identify such regional changes? Describe the reasoning for the approach you select, as well as the pros and cons.
4. You then perform an experiment to identify proteins that interact with MYB in the putative regulatory region that is specific to case samples. One of the proteins you identify is CBP – what would be the mechanistic sense given MYB’s known roles and the notion that the region is a putative enhancer?
5. The new putative enhancer region introduced in the case samples interacts with TAL1 gene region. Following on this observation, can you suggest the molecular sequence (step by step) leading to the condition of the case group? Further, now that you know the putative enhancer interacts with TAL1, what is the sub-category of the disease induced by the new MYB binding site?

**Part 5 – Networks (Hannah Carter):**

Submit answers to: pfiaux@ucsd.edu

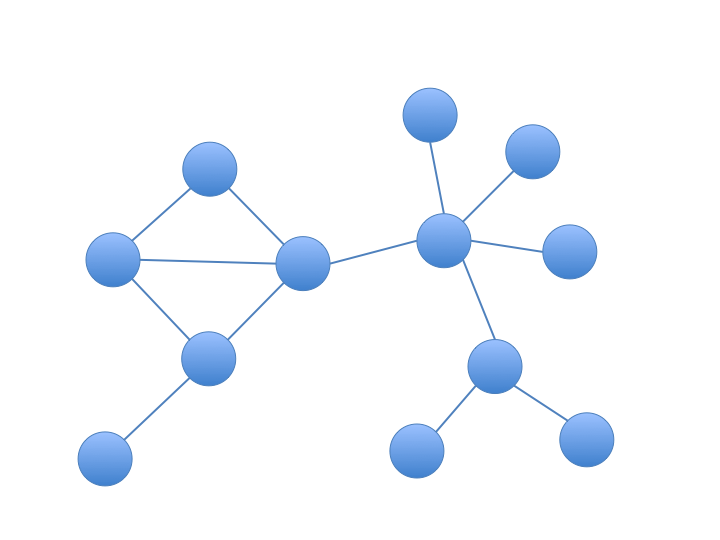
1

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?