BENG 183 Applied Genomic Technologies (Fall 2019)
Due: October 22, 2019 by 11:59pm

Homework 2

1. Intro to Precision Medicine

a. How is RNA detection and quantification useful in precision medicine? List at least two different techniques for the detection and/or quantification of RNA, and briefly describe their general principles.

b. A test has been developed to screen for cancer type A. Calculate the sensitivity and specificity of this test. Do you believe this test is a good screen for A?

	Has A	Does Not Have A
Screens Positive for A	34	40
Screens Negative for A	26	900

2. Intro to Machine Learning

RNA-seq data has been taken from several patients at a hospital, and the expression levels of 2 genes have been determined for each. Please complete the following questions **by hand**, although you are certainly welcome to use a software package to check your answer.

Patient	Gene 1 Expression	Gene 2 Expression
1	0.4	0.7
2	1.1	1.1
3	1.2	1.5
4	1.7	0.9
5	1.9	1.8

a. Perform hierarchical clustering on this gene expression data, using the Euclidean distance and single linkage method to calculate dissimilarity between clusters. Report results as a dendrogram and remember to show work!

b. Repeat the clustering of this data. Use the method of k-means, and start with points 1,4, and 5 as initial clusters. What are your final cluster assignments? If you start with random initial clusters, will you always get the same final cluster assignments?

3. ChIP-seq

In this question, we will explore how a genome browser can be used to look at ChIP-seq determined transcription factor binding sites. Go to the UCSC genome browser and select the GRCh37/hg19 assembly in the genomes tab. In the regulation tab below the genome browser, select the ENC TF Binding track. In the settings for this track, set the “ENCODE 3 TFBS” to show on the genome browser. By clicking on the “ENCODE 3 TFBS” hyperlink, you can see a table of the various transcription factors and cell types that have ChIP-seq data and can select only specific ones to show on the genome browser. Select your favorite transcription factor and cell type (e.g. SOX13 in HepG2 liver cell lines), and answer the following questions using the genome browser:

a. List the cell-type and transcription factor chosen and provide a brief description of them both (such as what this transcription factor is known to regulate, and what type of cells these are).
b. Find at least 2 different gene binding sites for your transcription factor. Attach images from the genome browser which show these binding sites. List the genes bound, and briefly describe the function of the gene (if known).

4. Genome Interaction Techniques I

a. Describe the various levels of genome folding in a eukaryotic nucleus.

b. What information regarding the chromosome do we gain from the Hi-C technique? Briefly describe the protocol used.

c. Describe the differences between the equilibrium and fractal globule models for chromatin architecture. Which is more appealing? Why?

5. Online tutorial 2: Raw Data QC and Cleaning

Please go through Tutorial 2 to QC check and clean the given sample sequences. Afterwards, complete the following tasks:

a. Attach a screenshot of your per sequence GC content for the 6h_1_fastqc and 2cells_1_fastqc fastQC reports.

b. From the fastp report, attach a screenshot of the 6hr sample duplication rate plot.

c. On the FastQC reports, the per base sequence content, per sequence GC content, and sequence duplication levels often show up as a “Fail” or a “Warning”. Are these sequences files of bad quality? If not, why would the program flag these categories as bad?