## Bioinformatics Assessment Brief

## Scenario

As a scientist working for the NHS on biomarker discovery in Nottingham. You were presented with unprocessed gene counts RNA sequencing data from 5 healthy colon tissues and 5 tumour colon tissues generated by (Kim, et al. 2014). By comparing cancerous and healthy samples, it is possible to identify differences in gene expression that can serve as biomarkers for colon cancer and as potential targets for new therapies.

**What you need to Do:**

Using your bioinformatic skills, you need to analyse this data to identify differentially expressed genes, ascertain the genomic locations of the top genes, and their functions and predict protein structure for the top differentially expressed genes.

* Please submit your report to the NOW DropBox as a single document in any commonly used file format, such as .docx or .pdf. Ensure that images and figures are embedded within the document rather than submitted as separate attachments. **Hard copies are not required.**
* Please answer the questions in the same order in which they have been presented below, using the same question numbering scheme.
* Pay attention to the sentence/word limits on some of the questions. If you exceed these limits, the rest of your response will not be considered towards your final mark for that question.

## Data

For this assessment, you will receive **one data file**:

Your dataset will be sent to you via email. Please download it and use ONLY this data for your assessment.

**Each student will receive a unique dataset**, meaning your results may differ from those of your peers. Be aware that collusion, including the sharing or copying of work, is strictly prohibited.

## Assessment questions

### Question 1

Now that you have your data, pre-process your data using the iDEP online tool (http://bioinformatics.sdstate.edu/idep11/). As a reminder, apply the parameters below for the filter:

* Genes with minimal 1.5 counts per million (CPM) in at least 5 samples
* Transform counts data using the EdgeR approach (CPM+c), considering 1 as pseudo count c.
* Treat the missing values as 0

 Use your result to answer the following questions (**a**) to (**d**).

1. How many genes remain in the expression matrix **after** normalization and filtering?
2. Create a plot of total RNA read counts for all samples, and import this in your report.
3. Explain what the plot result means **(Max 50 Words**)
4. Evaluate whether there are any patterns of similarity between your samples by using a Principal Component Analysis (PCA) plot (**Max 50 words**). Remember to import your plot in your report.

### Question 2

The next step is to perform a differential expression analysis by comparing cancer vs healthy samples, using the DESeq method with these parameters: FDR correction cutoff 0.05 and minimum fold-change 1.5. Use your result to answer the following questions (2a) to (2g).

1. How many genes were UP-regulated and how many were DOWN-regulated?
2. Use a volcano plot to display the top 10 differentially expressed genes from your analysis and include this in your report.
3. List the top **3** **UP-regulated** genes including their symbol, log2FoldChange and p-adjusted value in a tabular format in your report.
4. Identify the chromosome in the genome where each of these genes is located. Add this information to your table (created above).
5. List the top **3** **DOWN-regulated** genes including their symbol, log2FoldChange, and p-adjusted value in a tabular format in your report.
6. Identify the chromosome in the genome where each of these genes is located. Add this information to your table (created above).
7. List one function for each of your identified top **3 UP-regulated genes**. Add this information to your table (created above). Remember to cite your source of information.

### Question 3

### BLAST the DNA sequence of the top 2 differentially expressed genes against an appropriate database. Based on the results:

### identify the top 2 different species with high homology to each of your gene sequences.

### In your answer state:

### the BLAST program and the database used,

### how conserved are your top 2 differentially expressed genes across these species?

### the criteria for selecting these species (Max 50 words).

###

**For questions 4 -6**

Download the protein sequences of the top 3 most DOWN-regulated **protein-coding genes** in FASTA format. Use these sequences to answer questions 4 through 6.

### Question 4

Submit **EACH** of the downloaded protein sequences to Proparam (<https://web.expasy.org/protparam/>)

1. How many amino acids are present in the sequence?
2. What is the Molecular weight of the sequence?
3. What is the theoretical pI for the protein?
4. How many Cys residues are present for each sequence?
5. How can the information you found from the above questions help with characterising the protein in a laboratory setting or other practical application? (**100 words max**.)

###

### Question 5

Submit **EACH** of the downloaded protein sequences to Phyre2 (<http://www.sbg.bio.ic.ac.uk/~phyre2/>) to calculate a comparative homology model.

1. What is the confidence for the top 3 models and their percentage of identity for each protein sequence queried?
2. Choose the top model and report the percentage of sequence coverage and the Template information for each protein.
3. Align the top model for each sequence, which secondary structure elements can you recognise, and which one is dominant?

###

### Question 6

Open the PDB file for **ONE** of the top models you received from the Phyre2 server via email, using your preferred molecular visualisation software (e.g., PyMOL, 3D Protein Imaging, or others). Create a graphical representation of the protein, choosing an orientation and style that best highlights its features. You may include up to three images within the confines of half an A4 page. Select any file format that suits your needs, ensuring the image is large enough to clearly show the details you wish to highlight. If you aim to emphasise a specific feature of the protein, ensure it is clearly labeled for easy identification.

For any queries contact: adeolu.adewoye@ntu.ac.uk

### Good luck!