# Tempus Computational Biology

## Coding Challenge - RNA internship

Thank you for applying to Tempus. We're excited to continue the interview process through this coding challenge. RNA biomarker research in cancer often focuses on subtyping patients, then identifying correlates of those subtypes. In this coding challenge, you will complete a project in this area to get an idea of what it is like to work on our team. We expect the code challenge to take about half a day, and the reading assignment to take half a day. Please provide your code used to determine your answers along with a write up within a week.
The test includes:

1. A dataset of RNA expression (read counts) from tumor or adjacent normal tissue. Each row is a gene, and each column is a patient (data.csv).
2. 3 scientific articles about characterizing cancer.

Although the test is self contained, feel free to add any insight, and/or other data that you think is suitable.

Please answer the following questions to the best of your ability. Some of the questions do not have a definitive answer, and rather are designed for you to showcase your approach to problem solving.

Question 1: From the dataset, identify any outlier patients and describe the criteria used to identify them.,
Action: Remove outliers to obtain a cleaned up dataset of RNA counts.
Action: Calculate the most differentially expressed genes between tumors and normals.

Question 2: Which ten genes are the most upregulated in tumors, and which are the most downregulated?

Question 3: Identify if the correct tumor type is breast or sarcoma.
You can use the provided references from TCGA (TCGA-BRCA.pdf and TCGA-SARC.pdf) to help you determine tumor type. Alternatively, use any other approach to determine the tumor type.

Action: Normalize the RNA counts.
If you have never normalized RNA data, you can calculate log(count_ij + 1) where i is a row, and j is a column. Feel free to use any other RNA normalization method you prefer.

Question 4: How many tumor subtypes can you identify?
Usually we use PCA/UMAP/NMF to embed log normalized RNA data in low dimensions. Feel free to use any other method.

Question 5: How many more samples would you need to identify one more tumor subtype?

*Contact Rafi Pelossof (rafi.pelossof@tempus.com ) with any questions regarding this challenge.*