Alvin Hendricks (58104407), Carter Glen (16453292), Charlie Lake (51659233)

Progress Report

Introduction

Our report will start with discussing our project's goal: to investigate what genes characterize the different stages of bladder urethral carcinoma and analyze the impact of these genes on survival. We will then provide our methods for analysing the data: we will first perform hierarchical clustering on the RNA-seq data to visualise any groups that may be present. PCA dimensionality reduction will be used to allow for K-means clustering, used to separate varying expression groups. We expect to find that each cluster will be dominated by a specific stage of cancer. Each cluster will then be associated with specific mutations, allowing us to discuss the morbidity of specific mutations relative to cancer stage. Next, we discuss some expected challenges, such as hyperparameter selection for K-means clustering, and mapping the different datasets to each other. Finally, assignments for each project deliverable are given and a timeline is set.

Project Goal

The research question our team aims to investigate is: **"What genes characterize the various stages of bladder urethral carcinoma?"** Alongside this, we plan to look into the impacts of these mutations on survival rates, to better understand their impact. By identifying a list of specific genes that are common for each stage (1-4) of cancer, the expression of these genes may act as a powerful diagnostic tool to know not only the progression of cancer, but what treatments might be most appropriate. Understanding which genes are related to and possibly causing progression of cancer may allow for development of therapies that counteract the effects of these mutations.

Analysis Plan

In order to analyze which genes characterize various stages of bladder urethral carcinoma, we will need to identify which genes contribute most to varying expression. Given this list, we may be able to connect these genes back to patients with these mutations to find which ones are associated with each specific stage. So, we will primarily look at the gene expression of various patients based on their specific mutations. Once we hopefully identify the clusters, we'll analyze the survival rate of patients using the available data on patient life/death status and time from diagnosis. Our plan is summarised as follows:

1. Map the RNA counts to specific patients (samples) and specific mutations (EMSBL to Symbol mapping), and also map the mutations to specific patients. This will allow us to connect all the data and later associate gene expression with cancer stages.

2. Hierarchical clustering on RNA count (TPM) matrix, to organize data for visualization. Perform heatmap visualization of $Z(\log(Z.\ TPM + 1))$ counts, to identify groups of varying expression.

3. Run PCA for dimensionality reduction. This will let us reduce our dimensions to look at the key gene mutations that seem to affect expression.

4. Using trial and error by plotting gene expression against each PC pair, we will search for 2D PC vs. expression plots that provide clear separation of different groups. This will allow us to identify our clusters and estimate our number of clusters. We will count the number of groups found.

5. Run K-means clustering with k equal to the number of groups found in the previous step.

6. Run a count on each of our clusters, to see what proportion of each cluster is composed by each stage of cancer. Ideally, each cluster will be dominated by a specific stage of cancer, allowing us to identify each stage as characterized by a specific set of mutations .

7. Perform survival analysis on each of our clusters to see the impact of groups of gene expression.

8. Associate stages with specific mutations and discuss how they may affect mortality using our survival curves.

Challenges

A few challenges we anticipate encountering:

- Mapping the different data sets to each other might be difficult as there are large sections of IDs that overlap. For example, there are duplicate Symbols (due to multiple patients) that map to a single ENSEMBL ID in the RNA-Seq matrix. This should be feasible but uncertain as it has not been tried yet.
- Ideally, we would like to find 4 clusters with little overlap, with each one mapping to a specific stage of cancer. However, there may be multiple subtypes of each cancer stage (i.e., there may be multiple combinations of mutations that may characterize a specific cancer stage). So, we will have to plot the variation in counts against different PCs to find clear separation in our step 4, and to see how many groups exist. If it is found that more than 4 clusters exist, we will also need to analyze and determine how these clusters are distinct from each other, as it will not be fully dependent on the stage of cancer.

Timeline and Duties

We have identified regular times where we will meet to work on the project deliverables. Therefore, the majority of the tasks listed below will be completed together, with all group members present.

Week 1 (Nov 15th - Nov 19th):

- Complete the progress report (everyone did this collectively)
- Sample to Patient, ENSEMBL to Symbol mapping (Alvin)

Week 2 (Nov 22nd - Nov 26th):

- Start Analysis Steps 2-6

Week 3 (Nov 29th - Dec 3rd):

- Complete analysis steps 2-6
- Start and complete analysis steps 7-8
- Plan report and presentation format and sections

Week 4 (Dec 6th - Dec 10th):

- Complete the report and presentation